

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE  
**CERN** EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

## **2013 CERN–Latin-American School of High-Energy Physics**

Arequipa, Peru  
6 – 19 March 2013

**Proceedings**

Editors: M. Mulders  
G. Perez


ISBN 978-92-9083-412-0

ISSN 0531-4283

DOI <http://dx.doi.org/10.5170/CERN-2015-001>

Available online at <http://cds.cern.ch/>

Copyright © CERN, 2015

 Creative Commons Attribution 4.0

Knowledge transfer is an integral part of CERN's mission.

CERN publishes this report Open Access under the Creative Commons Attribution 4.0 license (<http://creativecommons.org/licenses/by/4.0/>) in order to permit its wide dissemination and use. The submission of a contribution to a CERN Yellow Report shall be deemed to constitute the contributor's agreement to this copyright and license statement. Contributors are requested to obtain any clearances that may be necessary for this purpose.

This report is indexed in: CERN Document Server (CDS), INSPIRE, Scopus.

This report should be cited as:

Proceedings of the 2013 CERN–Latin-American School of High-Energy Physics, Arequipa, Peru, 6–19 March 2013, edited by M. Mulders and G. Perez, CERN-2015-001 (CERN, Geneva, 2015), <http://dx.doi.org/10.5170/CERN-2015-001>

A contribution in this report should be cited as:

[Author name(s)], in Proceedings of the 2013 CERN–Latin-American School of High-Energy Physics, Arequipa, Peru, 6–19 March 2013, edited by M. Mulders and G. Perez, CERN-2015-001 (CERN, Geneva, 2015), pp. [first page]–[last page], <http://dx.doi.org/10.5170/CERN-2015-001>. [first page]

## **Abstract**

The CERN–Latin-American School of High-Energy Physics is intended to give young physicists an introduction to the theoretical aspects of recent advances in elementary particle physics. These proceedings contain lecture notes on the Standard Model of electroweak interactions, quantum chromodynamics, flavour physics, quantum chromodynamics under extreme conditions, cosmic-ray physics, cosmology, recent highlights of LHC results, practical statistics for particle physicists and a short introduction to the principles of particle physics instrumentation.



## Preface

The seventh School in the series of Latin-American Schools of High-Energy Physics took place from 6 to 19 March 2013 in Arequipa, Peru. It was organized by CERN with the support of local colleagues from several universities in Peru (PUCP, UNI and UNSA), with PUCP playing a leading role.

The School received financial support from: CERN; CIEMAT, Spain; RENAFAE, Brazil; and PUCP in Peru. Our sincere thanks go to all of these sponsors for making it possible to organize the School with many young participants from Latin-American countries who otherwise would not have been able to attend.

The School was hosted in the comfortable Estelar Hotel El Lago on the outskirts of the city of Arequipa. We are indebted to the hotel and its friendly staff for their help in making the event such a success. In particular, we would like to mention the hotel's general manager, Hugo Avila, who helped us greatly in preparing the School as well as during the event itself.

Professor Alberto Gago from PUCP acted as local director for the School, assisted by members of the local organising committee. We are extremely grateful to Alberto and his colleagues for their excellent work in organizing the School and for creating such a wonderful atmosphere for the participants. We would also like to mention the team from the physics department of the local university, UNSA, especially David Pacheco and Rolando Perca who helped with numerous practical arrangements.

Sixty-five students of 18 different nationalities attended the School. Following the tradition of the School the students shared twin rooms mixing nationalities, and in particular the Europeans mixed with Latin Americans.

The 11 lecturers came from Europe, Israel, Latin America and the USA. The lectures, which were given in English, were complemented by daily discussion sessions led by five physicists coming from Latin America. The lectures and the discussion sessions were all held using the conference facilities of the hotel. The students displayed their own research work in the form of posters in a special evening session during the first week. The posters were left on display until the end of the School. The students from each discussion group also performed a project, studying in detail the analysis of a published paper from an LHC experiment. A representative of each group presented a brief summary talk during a special evening session during the second week of the School.

Our thanks are due to the lecturers and discussion leaders for their active participation in the School and for making the scientific programme so stimulating. The students who in turn manifested their good spirits during two intense weeks undoubtedly appreciated their personal contributions in answering questions and explaining points of theory.

We are very grateful to H el ene Haller and Kate Ross, the Administrators for the CERN Schools of Physics, for their efforts in the lengthy preparations for the School and during the event itself. Their efficient work, friendly attitude, and continuous care of the participants and their needs were highly appreciated.

The participants will certainly remember the two interesting excursions, an afternoon tour of the city of Arequipa, and, particularly, a spectacular full-day excursion to the Colca Canyon for many of the participants, or to the Pacific coast for the others. They also greatly appreciated the excellent social and leisure programme, including horse riding and evenings spent together in the hotel, as well as the farewell party on the last night.

The success of the School was to a large extent due to the students themselves. Their poster session and group projects were very well prepared and highly appreciated, and throughout the School they participated actively during the lectures, in the discussion sessions, and in the different activities and excursions.

Nick Ellis  
(On behalf of the Organizing Committee)



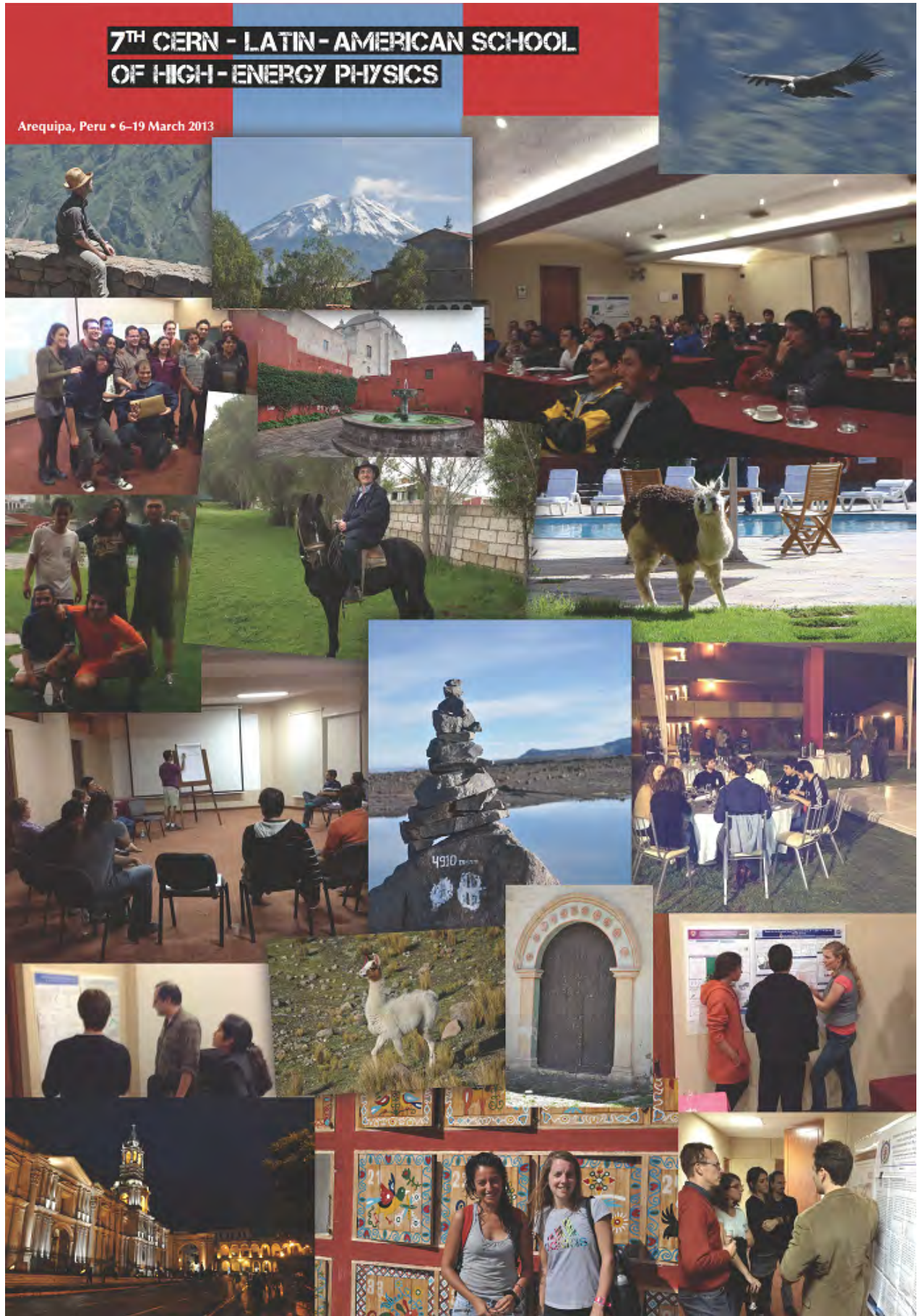
## People in the photograph

No.	Name	No.	Name	No.	Name	No.	Name		
1	Jose Alonso Carpio	21	Sony Martins	41	Anthony Canazas	61	Rolando Perca	81	Rosana Martinez Turtos
2	Oscar Vargas	22	Karim Massri	42	Alexander Parada	62	Oscar Gonzalez Lopez	82	Cecilia Jarne
3	Matthew Kenzie	23	Pedro Amao	43	Gabriela Cerqueira Gomes	63	Roger Naranjo	83	Nick Ellis
4	Mauricio Hippert	24	Edgar Valencia	44	Lorenzo Diaz Cruz	64	Juan Pablo Velásquez	84	Kate Ross
5	Simao Silva	25	Alessandro Manfredini	45	Bernabe Mejia	65	Estefania Coluccio Leskow	85	Alberto Gago
6	Gibraham Napoles	26	David Delepine	46	Andrés Melo	66	Maria Jose Bustamante	86	Martijn Mulders
7	Silvestre Romano	27	Luis Guillen	47	Orjan Dale	67	Alfonso Zerwekh	87	Jilberto Zamora Saa
8	Alex Dafinca	28	Frank Coronado	48	Lars Pedersen	68	Margot Delgado de la Flor	88	Jose Garcia
9	Duncan Leggat	29	Dennis Zavaleta	49	Alejandro Gomez	69	Adriano Sampieri	89	Canela (Alpaca)
10	Samantha Dooling	30	Edgar Carrera	50	Luis Alvarez Gaume	70	Yosef Nir		
11	David Hall	31	Pedro Malta	51	Pablo Jacome	71	Cynthia Vizcarra		
12	Jhovanny Mejia	32	Mauricio Suárez	52	Nicolas Mileo	72	Ivonne Maldonado		
13	Denis Robertson	33	Alberto Hernandez	53	Bruno Daniel	73	Jose Vega		
14	Jose La Rosa	34	Guillermo Palacio	54	Francisco Alonso	74	Jose-Carlos Jimenez		
15	Alejandro de la Puente	35	Lucas Cavalcanti	55	Alexander Arguello Quiroga	75	Carmen Araujo		
16	Gonzalo Diaz	36	Thamys Abrahao	56	Karoline Selbach	76	Janeth Valverde		
17	Miguel García	37	Vanessa Theodoro	57	Javier Solano	77	Marcela González		
18	Josue Molina	38	Hernan Castillo	58	Teofilo Vargas	78	Sabrina Sacerdoti		
19	Alan Gilberto Chavez	39	Fabio Maltoni	59	Joel Jones-Pérez	79	Giovanna Cottin		
20	Glauber Sampaio Dos Santos	40	Saúl Panibra	60	David Pacheco	80	Josefina Alconada		





# PHOTOGRAPHS (MONTAGE)





# Contents

Preface	
<i>N. Ellis</i> .....	v
Photograph of participants .....	vi
Photographs (montage) .....	ix
Introductory Lectures on Quantum Field Theory	
<i>L. Álvarez-Gaumé and M.A. Vázquez-Mozo</i> .....	1
Basics of QCD for the LHC: Higgs production as a case study	
<i>F. Maltoni</i> .....	95
Flavour Physics and CP Violation	
<i>Y. Nir</i> .....	123
QCD under extreme conditions: an informal discussion	
<i>E.S. Fraga</i> .....	157
Ultra-High-Energy Cosmic Rays	
<i>G.T. Dova</i> .....	169
LHC Results Highlights	
<i>O. González</i> .....	191
Particle Physics Instrumentation	
<i>W. Riegler</i> .....	235
Practical Statistics for the LHC	
<i>K. Cranmer</i> .....	247
Cosmology	
<i>J. García-Bellido</i> .....	291
Organizing Committee .....	369
Local Organizing Committee .....	369
List of Lecturers .....	369
List of Discussion Leaders .....	369
List of Students .....	370
List of Posters .....	371



# Introductory Lectures on Quantum Field Theory

*L. Álvarez-Gaumé<sup>a</sup> and M. A. Vázquez-Mozo<sup>b</sup>*

<sup>a</sup> CERN, Geneva, Switzerland

<sup>b</sup> Universidad de Salamanca, Salamanca, Spain

## Abstract

In these lectures we present a few topics in quantum field theory in detail. Some of them are conceptual and some more practical. They have been selected because they appear frequently in current applications to particle physics and string theory.

## 1 Introduction

These notes summarize lectures presented at the 2005 CERN-CLAF School in Malargüe (Argentina), the 2009 CERN-CLAF School in Medellín (Colombia), the 2011 CERN-CLAF School in Natal (Brazil), the 2012 Asia-Europe-Pacific School of High Energy Physics in Fukuoka (Japan), and the 2013 CERN–Latin-American School of High-Energy Physics in Arequipa (Peru). The audience in all occasions was composed to a large extent by students in experimental High Energy Physics with an important minority of theorists. In nearly ten hours it is quite difficult to give a reasonable introduction to a subject as vast as quantum field theory. For this reason the lectures were intended to provide a review of those parts of the subject to be used later by other lecturers. Although a cursory acquaintance with the subject of quantum field theory is helpful, the only requirement to follow the lectures is a working knowledge of Quantum Mechanics and Special Relativity.

The guiding principle in choosing the topics presented (apart to serve as introductions to later courses) was to present some basic aspects of the theory that present conceptual subtleties. Those topics one often is uncomfortable with after a first introduction to the subject. Among them we have selected:

- The need to introduce quantum fields, with the great complexity this implies.
- Quantization of gauge theories and the rôle of topology in quantum phenomena. We have included a brief study of the Aharonov-Bohm effect and Dirac's explanation of the quantization of the electric charge in terms of magnetic monopoles.
- Quantum aspects of global and gauge symmetries and their breaking.
- Anomalies.
- The physical idea behind the process of renormalization of quantum field theories.
- Some more specialized topics, like the creation of particles by classical fields and the very basics of supersymmetry.

These notes have been written following closely the original presentation, with numerous clarifications. Sometimes the treatment given to some subjects has been extended, in particular the discussion of the Casimir effect and particle creation by classical backgrounds. Since no group theory was assumed, we have included an Appendix with a review of the basic concepts.

By lack of space and purpose, few proofs have been included. Instead, very often we illustrate a concept or property by describing a physical situation where it arises. A very much expanded version of these lectures, following the same philosophy but including many other topics, has appeared in book form in [1]. For full details and proofs we refer the reader to the many textbooks in the subject, and in particular in the ones provided in the bibliography [2–11]. Specially modern presentations, very much in the spirit of these lectures, can be found in references [5, 6, 10, 11]. We should nevertheless warn the reader that we have been a bit cavalier about references. Our aim has been to provide mostly a (not exhaustive) list of reference for further reading. We apologize to those authors who feel misrepresented.

### A note about notation

Before starting it is convenient to review the notation used. Through these notes we will be using the metric  $\eta_{\mu\nu} = \text{diag}(1, -1, -1, -1)$ . Derivatives with respect to the four-vector  $x^\mu = (ct, \vec{x})$  will be denoted by the shorthand

$$\partial_\mu \equiv \frac{\partial}{\partial x^\mu} = \left( \frac{1}{c} \frac{\partial}{\partial t}, \vec{\nabla} \right). \quad (1)$$

As usual space-time indices will be labelled by Greek letters ( $\mu, \nu, \dots = 0, 1, 2, 3$ ) while Latin indices will be used for spatial directions ( $i, j, \dots = 1, 2, 3$ ). In many expressions we will use the notation  $\sigma^\mu = (\mathbf{1}, \sigma^i)$  where  $\sigma^i$  are the Pauli matrices

$$\sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (2)$$

Sometimes we use of the Feynman's slash notation  $\not{x} = \gamma^\mu a_\mu$ . Finally, unless stated otherwise, we work in natural units  $\hbar = c = 1$ .

## 2 Why do we need quantum field theory after all?

In spite of the impressive success of Quantum Mechanics in describing atomic physics, it was immediately clear after its formulation that its relativistic extension was not free of difficulties. These problems were clear already to Schrödinger, whose first guess for a wave equation of a free relativistic particle was the Klein-Gordon equation

$$\left( \frac{\partial^2}{\partial t^2} - \nabla^2 + m^2 \right) \psi(t, \vec{x}) = 0. \quad (3)$$

This equation follows directly from the relativistic “mass-shell” identity  $E^2 = \vec{p}^2 + m^2$  using the correspondence principle

$$\begin{aligned} E &\rightarrow i \frac{\partial}{\partial t}, \\ \vec{p} &\rightarrow -i \vec{\nabla}. \end{aligned} \quad (4)$$

Plane wave solutions to the wave equation (3) are readily obtained

$$\psi(t, \vec{x}) = e^{-ip_\mu x^\mu} = e^{-iEt + i\vec{p}\cdot\vec{x}} \quad \text{with} \quad E = \pm\omega_p \equiv \pm\sqrt{\vec{p}^2 + m^2}. \quad (5)$$

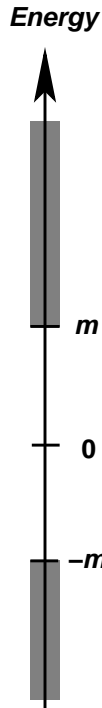
In order to have a complete basis of functions, one must include plane wave with both  $E > 0$  and  $E < 0$ . This implies that given the conserved current

$$j_\mu = \frac{i}{2} \left( \psi^* \partial_\mu \psi - \partial_\mu \psi^* \psi \right), \quad (6)$$

its time-component is  $j^0 = E$  and therefore does not define a positive-definite probability density.

A complete, properly normalized, continuous basis of solutions of the Klein-Gordon equation (3) labelled by the momentum  $\vec{p}$  can be defined as

$$\begin{aligned} f_p(t, \vec{x}) &= \frac{1}{(2\pi)^{\frac{3}{2}} \sqrt{2\omega_p}} e^{-i\omega_p t + i\vec{p}\cdot\vec{x}}, \\ f_{-p}(t, \vec{x}) &= \frac{1}{(2\pi)^{\frac{3}{2}} \sqrt{2\omega_p}} e^{i\omega_p t - i\vec{p}\cdot\vec{x}}. \end{aligned} \quad (7)$$



**Fig. 1:** Spectrum of the Klein-Gordon wave equation

Given the inner product

$$\langle \psi_1 | \psi_2 \rangle = i \int d^3x \left( \psi_1^* \partial_0 \psi_2 - \partial_0 \psi_1^* \psi_2 \right)$$

the states (7) form an orthonormal basis

$$\langle f_p | f_{p'} \rangle = \delta(\vec{p} - \vec{p}'),$$

$$\langle f_{-p} | f_{-p'} \rangle = -\delta(\vec{p} - \vec{p}'), \quad (8)$$

$$\langle f_p | f_{-p'} \rangle = 0. \quad (9)$$

The wave functions  $f_p(t, x)$  describes states with momentum  $\vec{p}$  and energy given by  $\omega_p = \sqrt{\vec{p}^2 + m^2}$ . On the other hand, the states  $|f_{-p}\rangle$  not only have a negative scalar product but they actually correspond to negative energy states

$$i\partial_0 f_{-p}(t, \vec{x}) = -\sqrt{\vec{p}^2 + m^2} f_{-p}(t, \vec{x}). \quad (10)$$

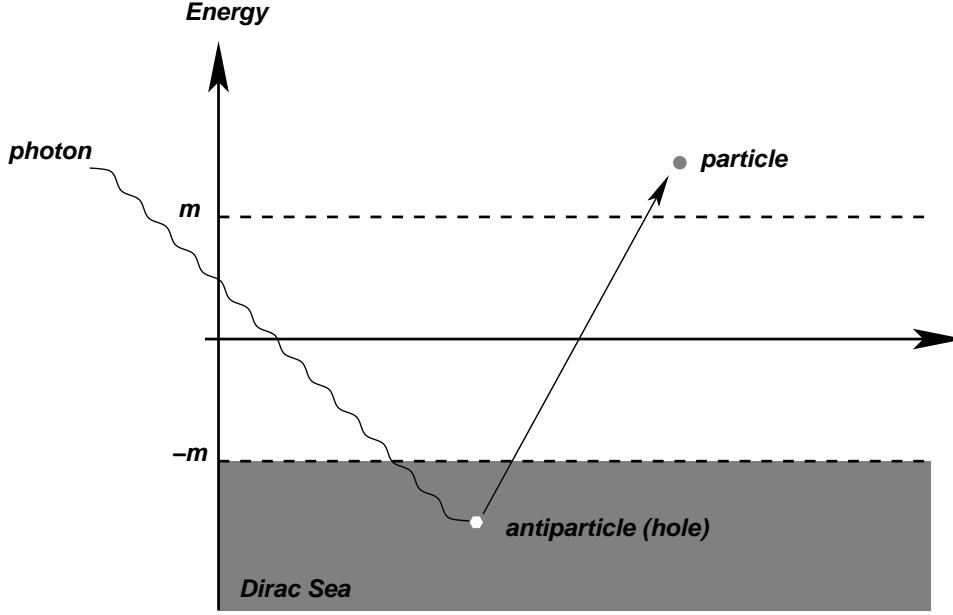
Therefore the energy spectrum of the theory satisfies  $|E| > m$  and is unbounded from below (see Fig. 1). Although in a case of a free theory the absence of a ground state is not necessarily a fatal problem, once the theory is coupled to the electromagnetic field this is the source of all kinds of disasters, since nothing can prevent the decay of any state by emission of electromagnetic radiation.

The problem of the instability of the “first-quantized” relativistic wave equation can be heuristically tackled in the case of spin- $\frac{1}{2}$  particles, described by the Dirac equation

$$\left( -i\beta \frac{\partial}{\partial t} + \vec{\alpha} \cdot \vec{\nabla} - m \right) \psi(t, \vec{x}) = 0, \quad (11)$$

where  $\vec{\alpha}$  and  $\beta$  are  $4 \times 4$  matrices

$$\alpha^i = \begin{pmatrix} 0 & i\sigma^i \\ -i\sigma^i & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}, \quad (12)$$



**Fig. 2:** Creation of a particle-antiparticle pair in the Dirac sea picture

with  $\sigma^i$  the Pauli matrices, and the wave function  $\psi(t, \vec{x})$  has four components. The wave equation (11) can be thought of as a kind of “square root” of the Klein-Gordon equation (3), since the latter can be obtained as

$$\left(-i\beta\frac{\partial}{\partial t} + \vec{\alpha} \cdot \vec{\nabla} - m\right)^\dagger \left(-i\beta\frac{\partial}{\partial t} + \vec{\alpha} \cdot \vec{\nabla} - m\right) \psi(t, \vec{x}) = \left(\frac{\partial^2}{\partial t^2} - \nabla^2 + m^2\right) \psi(t, \vec{x}). \quad (13)$$

An analysis of Eq. (11) along the lines of the one presented above for the Klein-Gordon equation leads again to the existence of negative energy states and a spectrum unbounded from below as in Fig. 1. Dirac, however, solved the instability problem by pointing out that now the particles are fermions and therefore they are subject to Pauli’s exclusion principle. Hence, each state in the spectrum can be occupied by at most one particle, so the states with  $E = m$  can be made stable if we assume that *all* the negative energy states are filled.

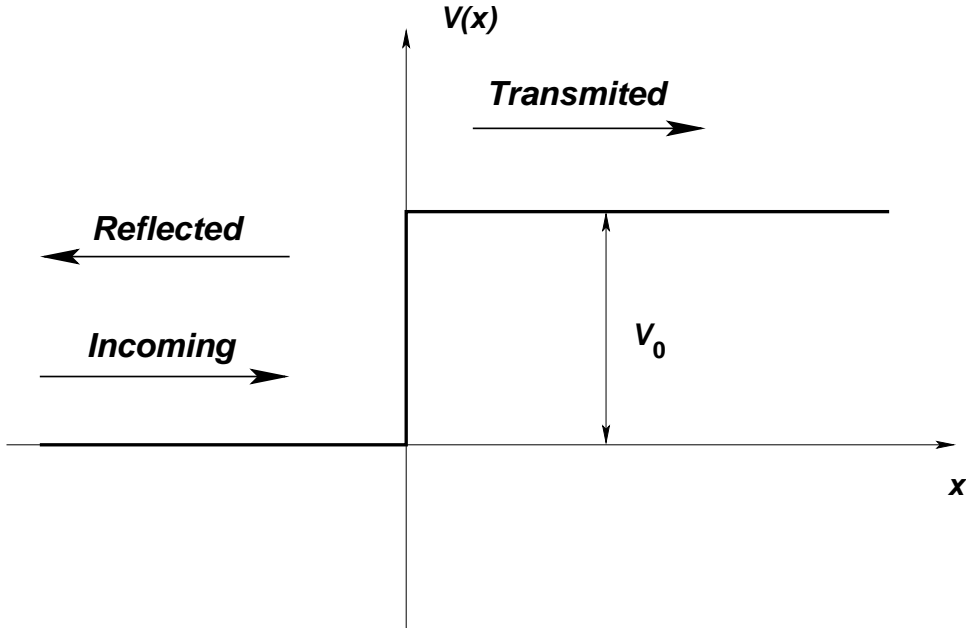
If Dirac’s idea restores the stability of the spectrum by introducing a stable vacuum where all negative energy states are occupied, the so-called Dirac sea, it also leads directly to the conclusion that a single-particle interpretation of the Dirac equation is not possible. Indeed, a photon with enough energy ( $E > 2m$ ) can excite one of the electrons filling the negative energy states, leaving behind a “hole” in the Dirac sea (see Fig. 2). This hole behaves as a particle with equal mass and opposite charge that is interpreted as a positron, so there is no escape to the conclusion that interactions will produce pairs particle-antiparticle out of the vacuum.

In spite of the success of the heuristic interpretation of negative energy states in the Dirac equation this is not the end of the story. In 1929 Oskar Klein stumbled into an apparent paradox when trying to describe the scattering of a relativistic electron by a square potential using Dirac’s wave equation [12] (for pedagogical reviews see [13, 14]). In order to capture the essence of the problem without entering into unnecessary complication we will study Klein’s paradox in the context of the Klein-Gordon equation.

Let us consider a square potential with height  $V_0 > 0$  of the type showed in Fig. 3. A solution to the wave equation in regions I and II is given by

$$\begin{aligned} \psi_I(t, x) &= e^{-iEt+ip_1x} + Re^{-iEt-ip_1x}, \\ \psi_{II}(t, x) &= Te^{-iEt+p_2x}, \end{aligned} \quad (14)$$





**Fig. 3:** Illustration of the Klein paradox.

where the mass-shell condition implies that

$$p_1 = \sqrt{E^2 - m^2}, \quad p_2 = \sqrt{(E - V_0)^2 - m^2}. \quad (15)$$

The constants  $R$  and  $T$  are computed by matching the two solutions across the boundary  $x = 0$ . The conditions  $\psi_I(t, 0) = \psi_{II}(t, 0)$  and  $\partial_x \psi_I(t, 0) = \partial_x \psi_{II}(t, 0)$  imply that

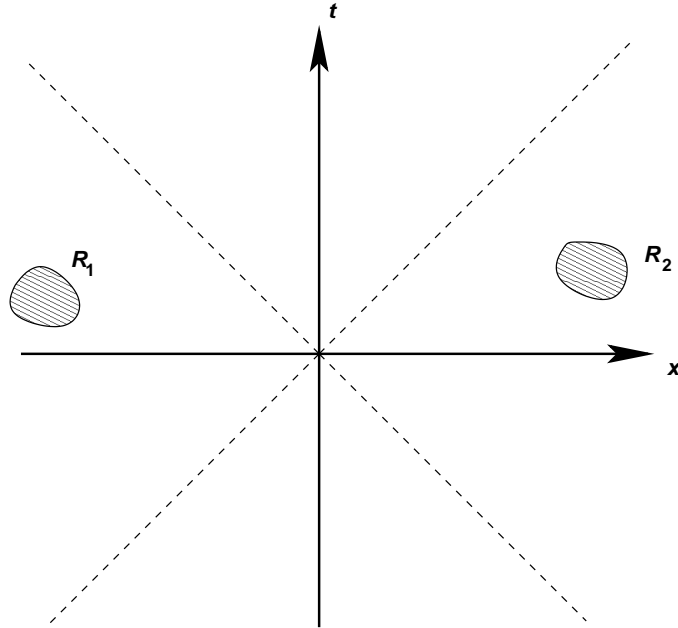
$$T = \frac{2p_1}{p_1 + p_2}, \quad R = \frac{p_1 - p_2}{p_1 + p_2}. \quad (16)$$

At first sight one would expect a behavior similar to the one encountered in the nonrelativistic case. If the kinetic energy is bigger than  $V_0$  both a transmitted and reflected wave are expected, whereas when the kinetic energy is smaller than  $V_0$  one only expect to find a reflected wave, the transmitted wave being exponentially damped within a distance of a Compton wavelength inside the barrier.

Indeed this is what happens if  $E - m > V_0$ . In this case both  $p_1$  and  $p_2$  are real and we have a partly reflected, and a partly transmitted wave. In the same way, if  $V_0 - 2m < E - m < V_0$  then  $p_2$  is imaginary and there is total reflection.

However, in the case when  $V_0 > 2m$  and the energy is in the range  $0 < E - m < V_0 - 2m$  a completely different situation arises. In this case one finds that both  $p_1$  and  $p_2$  are real and therefore the incoming wave function is partially reflected and partially transmitted across the barrier. This is a shocking result, since it implies that there is a nonvanishing probability of finding the particle at any point across the barrier with negative kinetic energy ( $E - m - V_0 < 0$ )! This weird result is known as Klein's paradox.

As with the negative energy states, the Klein paradox results from our insistence in giving a single-particle interpretation to the relativistic wave function. Actually, a multiparticle analysis of the paradox [13] shows that what happens when  $0 < E - m < V_0 - 2m$  is that the reflection of the incoming particle by the barrier is accompanied by the creation of pairs particle-antiparticle out of the energy of the barrier (notice that for this to happen it is required that  $V_0 > 2m$ , the threshold for the creation of a particle-antiparticle pair).



**Fig. 4:** Two regions  $R_1, R_2$  that are causally disconnected.

Actually, this particle creation can be understood by noticing that the sudden potential step in Fig. 3 localizes the incoming particle with mass  $m$  in distances smaller than its Compton wavelength  $\lambda = \frac{1}{m}$ . This can be seen by replacing the square potential by another one where the potential varies smoothly from 0 to  $V_0 > 2m$  in distances scales larger than  $1/m$ . This case was worked out by Sauter shortly after Klein pointed out the paradox [15]. He considered a situation where the regions with  $V = 0$  and  $V = V_0$  are connected by a region of length  $d$  with a linear potential  $V(x) = \frac{V_0 x}{d}$ . When  $d > \frac{1}{m}$  he found that the transmission coefficient is exponentially small<sup>1</sup>.

The creation of particles is impossible to avoid whenever one tries to locate a particle of mass  $m$  within its Compton wavelength. Indeed, from Heisenberg uncertainty relation we find that if  $\Delta x \sim \frac{1}{m}$ , the fluctuations in the momentum will be of order  $\Delta p \sim m$  and fluctuations in the energy of order

$$\Delta E \sim m \quad (17)$$

can be expected. Therefore, in a relativistic theory, the fluctuations of the energy are enough to allow the creation of particles out of the vacuum. In the case of a spin- $\frac{1}{2}$  particle, the Dirac sea picture shows clearly how, when the energy fluctuations are of order  $m$ , electrons from the Dirac sea can be excited to positive energy states, thus creating electron-positron pairs.

It is possible to see how the multiparticle interpretation is forced upon us by relativistic invariance. In non-relativistic Quantum Mechanics observables are represented by self-adjoint operator that in the Heisenberg picture depend on time. Therefore measurements are localized in time but are global in space. The situation is radically different in the relativistic case. Because no signal can propagate faster than the speed of light, measurements have to be localized both in time and space. Causality demands then that two measurements carried out in causally-disconnected regions of space-time cannot interfere with each other. In mathematical terms this means that if  $\mathcal{O}_{R_1}$  and  $\mathcal{O}_{R_2}$  are the observables associated with two measurements localized in two causally-disconnected regions  $R_1, R_2$  (see Fig. 4), they satisfy

$$[\mathcal{O}_{R_1}, \mathcal{O}_{R_2}] = 0, \quad \text{if } (x_1 - x_2)^2 < 0, \text{ for all } x_1 \in R_1, x_2 \in R_2. \quad (18)$$

<sup>1</sup>In section (9.1) we will see how, in the case of the Dirac field, this exponential behavior can be associated with the creation of electron-positron pairs due to a constant electric field (Schwinger effect).

Hence, in a relativistic theory, the basic operators in the Heisenberg picture must depend on the space-time position  $x^\mu$ . Unlike the case in non-relativistic quantum mechanics, here the position  $\vec{x}$  is *not* an observable, but just a label, similarly to the case of time in ordinary quantum mechanics. Causality is then imposed microscopically by requiring

$$[\mathcal{O}(x), \mathcal{O}(y)] = 0, \quad \text{if } (x - y)^2 < 0. \quad (19)$$

A smeared operator  $\mathcal{O}_R$  over a space-time region  $R$  can then be defined as

$$\mathcal{O}_R = \int d^4x \mathcal{O}(x) f_R(x) \quad (20)$$

where  $f_R(x)$  is the characteristic function associated with  $R$ ,

$$f_R(x) = \begin{cases} 1 & x \in R \\ 0 & x \notin R \end{cases}. \quad (21)$$

Eq. (18) follows now from the microcausality condition (19).

Therefore, relativistic invariance forces the introduction of quantum fields. It is only when we insist in keeping a single-particle interpretation that we crash against causality violations. To illustrate the point, let us consider a single particle wave function  $\psi(t, \vec{x})$  that initially is localized in the position  $\vec{x} = 0$

$$\psi(0, \vec{x}) = \delta(\vec{x}). \quad (22)$$

Evolving this wave function using the Hamiltonian  $H = \sqrt{-\nabla^2 + m^2}$  we find that the wave function can be written as

$$\psi(t, \vec{x}) = e^{-it\sqrt{-\nabla^2 + m^2}} \delta(\vec{x}) = \int \frac{d^3k}{(2\pi)^3} e^{i\vec{k}\cdot\vec{x} - it\sqrt{k^2 + m^2}}. \quad (23)$$

Integrating over the angular variables, the wave function can be recast in the form

$$\psi(t, \vec{x}) = \frac{1}{2\pi^2|\vec{x}|} \int_{-\infty}^{\infty} k dk e^{ik|\vec{x}|} e^{-it\sqrt{k^2 + m^2}}. \quad (24)$$

The resulting integral can be evaluated using the complex integration contour  $C$  shown in Fig. 5. The result is that, for any  $t > 0$ , one finds that  $\psi(t, \vec{x}) \neq 0$  for any  $\vec{x}$ . If we insist in interpreting the wave function  $\psi(t, \vec{x})$  as the probability density of finding the particle at the location  $\vec{x}$  in the time  $t$  we find that the probability leaks out of the light cone, thus violating causality.

### 3 From classical to quantum fields

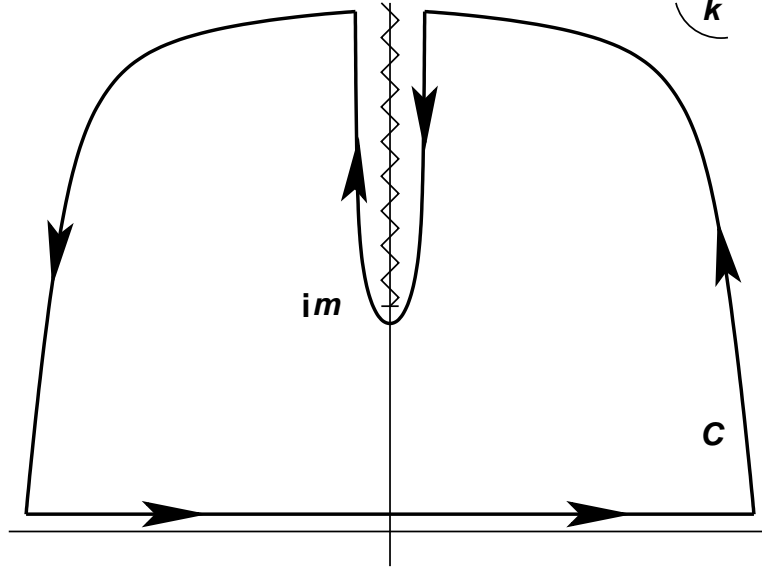
We have learned how the consistency of quantum mechanics with special relativity forces us to abandon the single-particle interpretation of the wave function. Instead we have to consider quantum fields whose elementary excitations are associated with particle states, as we will see below.

In any scattering experiment, the only information available to us is the set of quantum number associated with the set of free particles in the initial and final states. Ignoring for the moment other quantum numbers like spin and flavor, one-particle states are labelled by the three-momentum  $\vec{p}$  and span the single-particle Hilbert space  $\mathcal{H}_1$

$$|\vec{p}\rangle \in \mathcal{H}_1, \quad \langle \vec{p} | \vec{p}' \rangle = \delta(\vec{p} - \vec{p}'). \quad (25)$$

The states  $\{|\vec{p}\rangle\}$  form a basis of  $\mathcal{H}_1$  and therefore satisfy the closure relation

$$\int d^3p |\vec{p}\rangle \langle \vec{p}| = \mathbf{1} \quad (26)$$



**Fig. 5:** Complex contour  $C$  for the computation of the integral in Eq. (24).

The group of spatial rotations acts unitarily on the states  $|\vec{p}\rangle$ . This means that for every rotation  $R \in \text{SO}(3)$  there is a unitary operator  $\mathcal{U}(R)$  such that

$$\mathcal{U}(R)|\vec{p}\rangle = |R\vec{p}\rangle \quad (27)$$

where  $R\vec{p}$  represents the action of the rotation on the vector  $\vec{k}$ ,  $(R\vec{p})^i = R^i_j k^j$ . Using a spectral decomposition, the momentum operator  $\hat{P}^i$  can be written as

$$\hat{P}^i = \int d^3p |\vec{p}\rangle p^i \langle \vec{p}| \quad (28)$$

With the help of Eq. (27) it is straightforward to check that the momentum operator transforms as a vector under rotations:

$$\mathcal{U}(R)^{-1} \hat{P}^i \mathcal{U}(R) = \int d^3p |R^{-1}\vec{p}\rangle p^i \langle R^{-1}\vec{p}| = R^i_j \hat{P}^j, \quad (29)$$

where we have used that the integration measure is invariant under  $\text{SO}(3)$ .

Since, as we argued above, we are forced to deal with multiparticle states, it is convenient to introduce creation-annihilation operators associated with a single-particle state of momentum  $\vec{p}$

$$[a(\vec{p}), a^\dagger(\vec{p}')] = \delta(\vec{p} - \vec{p}'), \quad [a(\vec{p}), a(\vec{p}')] = [a^\dagger(\vec{p}), a^\dagger(\vec{p}')] = 0, \quad (30)$$

such that the state  $|\vec{p}\rangle$  is created out of the Fock space vacuum  $|0\rangle$  (normalized such that  $\langle 0|0\rangle = 1$ ) by the action of a creation operator  $a^\dagger(\vec{p})$

$$|\vec{p}\rangle = a^\dagger(\vec{p})|0\rangle, \quad a(\vec{p})|0\rangle = 0 \quad \forall \vec{p}. \quad (31)$$

Covariance under spatial rotations is all we need if we are interested in a nonrelativistic theory. However in a relativistic quantum field theory we must preserve more than  $\text{SO}(3)$ , actually we need the expressions to be covariant under the full Poincaré group  $\text{ISO}(1, 3)$  consisting in spatial rotations, boosts and space-time translations. Therefore, in order to build the Fock space of the theory we need two key ingredients: first an invariant normalization for the states, since we want a normalized state in

one reference frame to be normalized in any other inertial frame. And secondly a relativistic invariant integration measure in momentum space, so the spectral decomposition of operators is covariant under the full Poincaré group.

Let us begin with the invariant measure. Given an invariant function  $f(p)$  of the four-momentum  $p^\mu$  of a particle of mass  $m$  with positive energy  $p^0 > 0$ , there is an integration measure which is invariant under proper Lorentz transformations<sup>2</sup>

$$\int \frac{d^4 p}{(2\pi)^4} (2\pi) \delta(p^2 - m^2) \theta(p^0) f(p), \quad (32)$$

where  $\theta(x)$  represent the Heaviside step function. The integration over  $p^0$  can be easily done using the  $\delta$ -function identity

$$\delta[f(x)] = \sum_{x_i = \text{zeros of } f} \frac{1}{|f'(x_i)|} \delta(x - x_i), \quad (33)$$

which in our case implies that

$$\delta(p^2 - m^2) = \frac{1}{2p^0} \delta\left(p^0 - \sqrt{\vec{p}^2 + m^2}\right) + \frac{1}{2p^0} \delta\left(p^0 + \sqrt{\vec{p}^2 + m^2}\right). \quad (34)$$

The second term in the previous expression correspond to states with negative energy and therefore does not contribute to the integral. We can write then

$$\int \frac{d^4 p}{(2\pi)^4} (2\pi) \delta(p^2 - m^2) \theta(p^0) f(p) = \int \frac{d^3 p}{(2\pi)^3} \frac{1}{2\sqrt{\vec{p}^2 + m^2}} f\left(\sqrt{\vec{p}^2 + m^2}, \vec{p}\right). \quad (35)$$

Hence, the relativistic invariant measure is given by

$$\int \frac{d^3 p}{(2\pi)^3} \frac{1}{2\omega_p} \quad \text{with} \quad \omega_p \equiv \sqrt{\vec{p}^2 + m^2}. \quad (36)$$

Once we have an invariant measure the next step is to find an invariant normalization for the states. We work with a basis  $\{|p\rangle\}$  of eigenstates of the four-momentum operator  $\hat{P}^\mu$

$$\hat{P}^0 |p\rangle = \omega_p |p\rangle, \quad \hat{P}^i |p\rangle = p^i |p\rangle. \quad (37)$$

Since the states  $|p\rangle$  are eigenstates of the three-momentum operator we can express them in terms of the non-relativistic states  $|\vec{p}\rangle$  that we introduced in Eq. (25)

$$|p\rangle = N(\vec{p}) |\vec{p}\rangle \quad (38)$$

with  $N(\vec{p})$  a normalization to be determined now. The states  $\{|p\rangle\}$  form a complete basis, so they should satisfy the Lorentz invariant closure relation

$$\int \frac{d^4 p}{(2\pi)^4} (2\pi) \delta(p^2 - m^2) \theta(p^0) |p\rangle \langle p| = \mathbf{1} \quad (39)$$

At the same time, this closure relation can be expressed, using Eq. (38), in terms of the nonrelativistic basis of states  $\{|\vec{p}\rangle\}$  as

$$\int \frac{d^4 p}{(2\pi)^4} (2\pi) \delta(p^2 - m^2) \theta(p^0) |p\rangle \langle p| = \int \frac{d^3 p}{(2\pi)^3} \frac{1}{2\omega_p} |N(p)|^2 |\vec{p}\rangle \langle \vec{p}|. \quad (40)$$

<sup>2</sup>The factors of  $2\pi$  are introduced for later convenience.

Using now Eq. (28) for the nonrelativistic states, expression (39) follows provided

$$|N(\vec{p})|^2 = (2\pi)^3 (2\omega_p). \quad (41)$$

Taking the overall phase in Eq. (38) so that  $N(p)$  is real, we define the Lorentz invariant states  $|p\rangle$  as

$$|p\rangle = (2\pi)^{\frac{3}{2}} \sqrt{2\omega_p} |\vec{p}\rangle, \quad (42)$$

and given the normalization of  $|\vec{p}\rangle$  we find the normalization of the relativistic states to be

$$\langle p|p'\rangle = (2\pi)^3 (2\omega_p) \delta(\vec{p} - \vec{p}'). \quad (43)$$

Although not obvious at first sight, the previous normalization is Lorentz invariant. Although it is not difficult to show this in general, here we consider the simpler case of 1+1 dimensions where the two components  $(p^0, p^1)$  of the on-shell momentum can be parametrized in terms of a single hyperbolic angle  $\lambda$  as

$$p^0 = m \cosh \lambda, \quad p^1 = m \sinh \lambda. \quad (44)$$

Now, the combination  $2\omega_p \delta(p^1 - p^{1'})$  can be written as

$$2\omega_p \delta(p^1 - p^{1'}) = 2m \cosh \lambda \delta(m \sinh \lambda - m \sinh \lambda') = 2\delta(\lambda - \lambda'), \quad (45)$$

where we have made use of the property (33) of the  $\delta$ -function. Lorentz transformations in 1 + 1 dimensions are labelled by a parameter  $\xi \in \mathbb{R}$  and act on the momentum by shifting the hyperbolic angle  $\lambda \rightarrow \lambda + \xi$ . However, Eq. (45) is invariant under a common shift of  $\lambda$  and  $\lambda'$ , so the whole expression is obviously invariant under Lorentz transformations.

To summarize what we did so far, we have succeed in constructing a Lorentz covariant basis of states for the one-particle Hilbert space  $\mathcal{H}_1$ . The generators of the Poincaré group act on the states  $|p\rangle$  of the basis as

$$\widehat{P}^\mu |p\rangle = p^\mu |p\rangle, \quad \mathcal{U}(\Lambda) |p\rangle = |\Lambda^\mu{}_\nu p^\nu\rangle \equiv |\Lambda p\rangle \quad \text{with} \quad \Lambda \in \text{SO}(1, 3). \quad (46)$$

This is compatible with the Lorentz invariance of the normalization that we have checked above

$$\langle p|p'\rangle = \langle p|\mathcal{U}(\Lambda)^{-1}\mathcal{U}(\Lambda)|p'\rangle = \langle \Lambda p|\Lambda p'\rangle. \quad (47)$$

On  $\mathcal{H}_1$  the operator  $\widehat{P}^\mu$  admits the following spectral representation

$$\widehat{P}^\mu = \int \frac{d^3p}{(2\pi)^3} \frac{1}{2\omega_p} |p\rangle p^\mu \langle p|. \quad (48)$$

Using (47) and the fact that the measure is invariant under Lorentz transformation, one can easily show that  $\widehat{P}^\mu$  transform covariantly under  $\text{SO}(1, 3)$

$$\mathcal{U}(\Lambda)^{-1} \widehat{P}^\mu \mathcal{U}(\Lambda) = \int \frac{d^3p}{(2\pi)^3} \frac{1}{2\omega_p} |\Lambda^{-1}p\rangle p^\mu \langle \Lambda^{-1}p| = \Lambda^\mu{}_\nu \widehat{P}^\nu. \quad (49)$$

A set of covariant creation-annihilation operators can be constructed now in terms of the operators  $a(\vec{p})$ ,  $a^\dagger(\vec{p})$  introduced above

$$\alpha(\vec{p}) \equiv (2\pi)^{\frac{3}{2}} \sqrt{2\omega_p} a(\vec{p}), \quad \alpha^\dagger(\vec{p}) \equiv (2\pi)^{\frac{3}{2}} \sqrt{2\omega_p} a^\dagger(\vec{p}) \quad (50)$$

with the Lorentz invariant commutation relations

$$[\alpha(\vec{p}), \alpha^\dagger(\vec{p}')] = (2\pi)^3 (2\omega_p) \delta(\vec{p} - \vec{p}'),$$

$$[\alpha(\vec{p}), \alpha(\vec{p}')] = [\alpha^\dagger(\vec{p}), \alpha^\dagger(\vec{p}')] = 0. \quad (51)$$

Particle states are created by acting with any number of creation operators  $\alpha(\vec{p})$  on the Poincaré invariant vacuum state  $|0\rangle$  satisfying

$$\langle 0|0\rangle = 1, \quad \widehat{P}^\mu|0\rangle = 0, \quad \mathcal{U}(\Lambda)|0\rangle = |0\rangle, \quad \forall \Lambda \in \text{SO}(1, 3). \quad (52)$$

A general one-particle state  $|f\rangle \in \mathcal{H}_1$  can be then written as

$$|f\rangle = \int \frac{d^3p}{(2\pi)^3} \frac{1}{2\omega_p} f(\vec{p}) \alpha^\dagger(\vec{p})|0\rangle, \quad (53)$$

while a  $n$ -particle state  $|f\rangle \in \mathcal{H}_1^{\otimes n}$  can be expressed as

$$|f\rangle = \int \prod_{i=1}^n \frac{d^3p_i}{(2\pi)^3} \frac{1}{2\omega_{p_i}} f(\vec{p}_1, \dots, \vec{p}_n) \alpha^\dagger(\vec{p}_1) \dots \alpha^\dagger(\vec{p}_n)|0\rangle. \quad (54)$$

That this states are Lorentz invariant can be checked by noticing that from the definition of the creation-annihilation operators follows the transformation

$$\mathcal{U}(\Lambda)\alpha(\vec{p})\mathcal{U}(\Lambda)^\dagger = \alpha(\Lambda\vec{p}) \quad (55)$$

and the corresponding one for creation operators.

As we have argued above, the very fact that measurements have to be localized implies the necessity of introducing quantum fields. Here we will consider the simplest case of a scalar quantum field  $\phi(x)$  satisfying the following properties:

- **Hermiticity.**

$$\phi^\dagger(x) = \phi(x). \quad (56)$$

- **Microcausality.** Since measurements cannot interfere with each other when performed in causally disconnected points of space-time, the commutator of two fields have to vanish outside the relative lighth-cone

$$[\phi(x), \phi(y)] = 0, \quad (x - y)^2 < 0. \quad (57)$$

- **Translation invariance.**

$$e^{i\widehat{P}\cdot a}\phi(x)e^{-i\widehat{P}\cdot a} = \phi(x - a). \quad (58)$$

- **Lorentz invariance.**

$$\mathcal{U}(\Lambda)^\dagger\phi(x)\mathcal{U}(\Lambda) = \phi(\Lambda^{-1}x). \quad (59)$$

- **Linearity.** To simplify matters we will also assume that  $\phi(x)$  is linear in the creation-annihilation operators  $\alpha(\vec{p}), \alpha^\dagger(\vec{p})$

$$\phi(x) = \int \frac{d^3p}{(2\pi)^3} \frac{1}{2\omega_p} \left[ f(\vec{p}, x)\alpha(\vec{p}) + g(\vec{p}, x)\alpha^\dagger(\vec{p}) \right]. \quad (60)$$

Since  $\phi(x)$  should be hermitian we are forced to take  $f(\vec{p}, x)^* = g(\vec{p}, x)$ . Moreover,  $\phi(x)$  satisfies the equations of motion of a free scalar field,  $(\partial_\mu\partial^\mu + m^2)\phi(x) = 0$ , only if  $f(\vec{p}, x)$  is a complete basis of solutions of the Klein-Gordon equation. These considerations leads to the expansion

$$\phi(x) = \int \frac{d^3p}{(2\pi)^3} \frac{1}{2\omega_p} \left[ e^{-i\omega_p t + i\vec{p}\cdot\vec{x}} \alpha(\vec{p}) + e^{i\omega_p t - i\vec{p}\cdot\vec{x}} \alpha^\dagger(\vec{p}) \right]. \quad (61)$$

Given the expansion of the scalar field in terms of the creation-annihilation operators it can be checked that  $\phi(x)$  and  $\partial_t\phi(x)$  satisfy the equal-time canonical commutation relations

$$[\phi(t, \vec{x}), \partial_t\phi(t, \vec{y})] = i\delta(\vec{x} - \vec{y}) \quad (62)$$

The general commutator  $[\phi(x), \phi(y)]$  can be also computed to be

$$[\phi(x), \phi(x')] = i\Delta(x - x'). \quad (63)$$

The function  $\Delta(x - y)$  is given by

$$\begin{aligned} i\Delta(x - y) &= -\text{Im} \int \frac{d^3p}{(2\pi)^3} \frac{1}{2\omega_p} e^{-i\omega_p(t-t') + i\vec{p}\cdot(\vec{x}-\vec{x}')} \\ &= \int \frac{d^4p}{(2\pi)^4} (2\pi)\delta(p^2 - m^2)\varepsilon(p^0)e^{-ip\cdot(x-x')}, \end{aligned} \quad (64)$$

where  $\varepsilon(x)$  is defined as

$$\varepsilon(x) \equiv \theta(x) - \theta(-x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases}. \quad (65)$$

Using the last expression in Eq. (64) it is easy to show that  $i\Delta(x - x')$  vanishes when  $x$  and  $x'$  are space-like separated. Indeed, if  $(x - x')^2 < 0$  there is always a reference frame in which both events are simultaneous, and since  $i\Delta(x - x')$  is Lorentz invariant we can compute it in this reference frame. In this case  $t = t'$  and the exponential in the second line of (64) does not depend on  $p^0$ . Therefore, the integration over  $k^0$  gives

$$\begin{aligned} \int_{-\infty}^{\infty} dp^0 \varepsilon(p^0)\delta(p^2 - m^2) &= \int_{-\infty}^{\infty} dp^0 \left[ \frac{1}{2\omega_p} \varepsilon(p^0)\delta(p^0 - \omega_p) + \frac{1}{2\omega_p} \varepsilon(p^0)\delta(p^0 + \omega_p) \right] \\ &= \frac{1}{2\omega_p} - \frac{1}{2\omega_p} = 0. \end{aligned} \quad (66)$$

So we have concluded that  $i\Delta(x - x') = 0$  if  $(x - x')^2 < 0$ , as required by microcausality. Notice that the situation is completely different when  $(x - x')^2 \geq 0$ , since in this case the exponential depends on  $p^0$  and the integration over this component of the momentum does not vanish.

### 3.1 Canonical quantization

So far we have contented ourselves with requiring a number of properties to the quantum scalar field: existence of asymptotic states, locality, microcausality and relativistic invariance. With these only ingredients we have managed to go quite far. The previous can also be obtained using canonical quantization. One starts with a classical free scalar field theory in Hamiltonian formalism and obtains the quantum theory by replacing Poisson brackets by commutators. Since this quantization procedure is based on the use of the canonical formalism, which gives time a privileged rôle, it is important to check at the end of the calculation that the resulting quantum theory is Lorentz invariant. In the following we will briefly overview the canonical quantization of the Klein-Gordon scalar field.

The starting point is the action functional  $S[\phi(x)]$  which, in the case of a free real scalar field of mass  $m$  is given by

$$S[\phi(x)] \equiv \int d^4x \mathcal{L}(\phi, \partial_\mu\phi) = \frac{1}{2} \int d^4x (\partial_\mu\phi\partial^\mu\phi - m^2\phi^2). \quad (67)$$



The equations of motion are obtained, as usual, from the Euler-Lagrange equations

$$\partial_\mu \left[ \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \right] - \frac{\partial \mathcal{L}}{\partial \phi} = 0 \quad \Longrightarrow \quad (\partial_\mu \partial^\mu + m^2)\phi = 0. \quad (68)$$

The momentum canonically conjugated to the field  $\phi(x)$  is given by

$$\pi(x) \equiv \frac{\partial \mathcal{L}}{\partial(\partial_0 \phi)} = \frac{\partial \phi}{\partial t}. \quad (69)$$

In the Hamiltonian formalism the physical system is described not in terms of the generalized coordinates and their time derivatives but in terms of the generalized coordinates and their canonically conjugated momenta. This is achieved by a Legendre transformation after which the dynamics of the system is determined by the Hamiltonian function

$$H \equiv \int d^3x \left( \pi \frac{\partial \phi}{\partial t} - \mathcal{L} \right) = \frac{1}{2} \int d^3x \left[ \pi^2 + (\vec{\nabla} \phi)^2 + m^2 \right]. \quad (70)$$

The equations of motion can be written in terms of the Poisson brackets. Given two functional  $A[\phi, \pi]$ ,  $B[\phi, \pi]$  of the canonical variables

$$A[\phi, \pi] = \int d^3x \mathcal{A}(\phi, \pi), \quad B[\phi, \pi] = \int d^3x \mathcal{B}(\phi, \pi). \quad (71)$$

Their Poisson bracket is defined by

$$\{A, B\} \equiv \int d^3x \left[ \frac{\delta A}{\delta \phi} \frac{\delta B}{\delta \pi} - \frac{\delta A}{\delta \pi} \frac{\delta B}{\delta \phi} \right], \quad (72)$$

where  $\frac{\delta}{\delta \phi}$  denotes the functional derivative defined as

$$\frac{\delta A}{\delta \phi} \equiv \frac{\partial \mathcal{A}}{\partial \phi} - \partial_\mu \left[ \frac{\partial \mathcal{A}}{\partial(\partial_\mu \phi)} \right] \quad (73)$$

Then, the canonically conjugated fields satisfy the following equal time Poisson brackets

$$\begin{aligned} \{\phi(t, \vec{x}), \phi(t, \vec{x}')\} &= \{\pi(t, \vec{x}), \pi(t, \vec{x}')\} = 0, \\ \{\phi(t, \vec{x}), \pi(t, \vec{x}')\} &= \delta(\vec{x} - \vec{x}'). \end{aligned} \quad (74)$$

Canonical quantization proceeds now by replacing classical fields with operators and Poisson brackets with commutators according to the rule

$$i\{\cdot, \cdot\} \longrightarrow [\cdot, \cdot]. \quad (75)$$

In the case of the scalar field, a general solution of the field equations (68) can be obtained by working with the Fourier transform

$$(\partial_\mu \partial^\mu + m^2)\phi(x) = 0 \quad \Longrightarrow \quad (-p^2 + m^2)\tilde{\phi}(p) = 0, \quad (76)$$

whose general solution can be written as<sup>3</sup>

$$\phi(x) = \int \frac{d^4p}{(2\pi)^4} (2\pi) \delta(p^2 - m^2) \theta(p^0) [\alpha(p) e^{-ip \cdot x} + \alpha(p)^* e^{ip \cdot x}]$$

<sup>3</sup>In momentum space, the general solution to this equation is  $\tilde{\phi}(p) = f(p) \delta(p^2 - m^2)$ , with  $f(p)$  a completely general function of  $p^\mu$ . The solution in position space is obtained by inverse Fourier transform.

$$= \int \frac{d^3p}{(2\pi)^3} \frac{1}{2\omega_p} \left[ \alpha(\vec{p}) e^{-i\omega_p t + \vec{p} \cdot \vec{x}} + \alpha(\vec{p})^* e^{i\omega_p t - \vec{p} \cdot \vec{x}} \right] \quad (77)$$

and we have required  $\phi(x)$  to be real. The conjugate momentum is

$$\pi(x) = -\frac{i}{2} \int \frac{d^3p}{(2\pi)^3} \left[ \alpha(\vec{p}) e^{-i\omega_p t + \vec{p} \cdot \vec{x}} + \alpha(\vec{p})^* e^{i\omega_p t - \vec{p} \cdot \vec{x}} \right]. \quad (78)$$

Now  $\phi(x)$  and  $\pi(x)$  are promoted to operators by replacing the functions  $\alpha(\vec{p})$ ,  $\alpha(\vec{p})^*$  by the corresponding operators

$$\alpha(\vec{p}) \longrightarrow \hat{\alpha}(\vec{p}), \quad \alpha(\vec{p})^* \longrightarrow \hat{\alpha}^\dagger(\vec{p}). \quad (79)$$

Moreover, demanding  $[\phi(t, \vec{x}), \pi(t, \vec{x}')] = i\delta(\vec{x} - \vec{x}')$  forces the operators  $\hat{\alpha}(\vec{p})$ ,  $\hat{\alpha}(\vec{p})^\dagger$  to have the commutation relations found in Eq. (51). Therefore they are identified as a set of creation-annihilation operators creating states with well-defined momentum  $\vec{p}$  out of the vacuum  $|0\rangle$ . In the canonical quantization formalism the concept of particle appears as a result of the quantization of a classical field.

Knowing the expressions of  $\hat{\phi}$  and  $\hat{\pi}$  in terms of the creation-annihilation operators we can proceed to evaluate the Hamiltonian operator. After a simple calculation one arrives to the expression

$$\hat{H} = \int d^3p \left[ \omega_p \hat{\alpha}^\dagger(\vec{p}) \hat{\alpha}(\vec{p}) + \frac{1}{2} \omega_p \delta(\vec{0}) \right]. \quad (80)$$

The first term has a simple physical interpretation since  $\hat{\alpha}^\dagger(\vec{p}) \hat{\alpha}(\vec{p})$  is the number operator of particles with momentum  $\vec{p}$ . The second divergent term can be eliminated if we defined the normal-ordered Hamiltonian  $:\hat{H}:$  with the vacuum energy subtracted

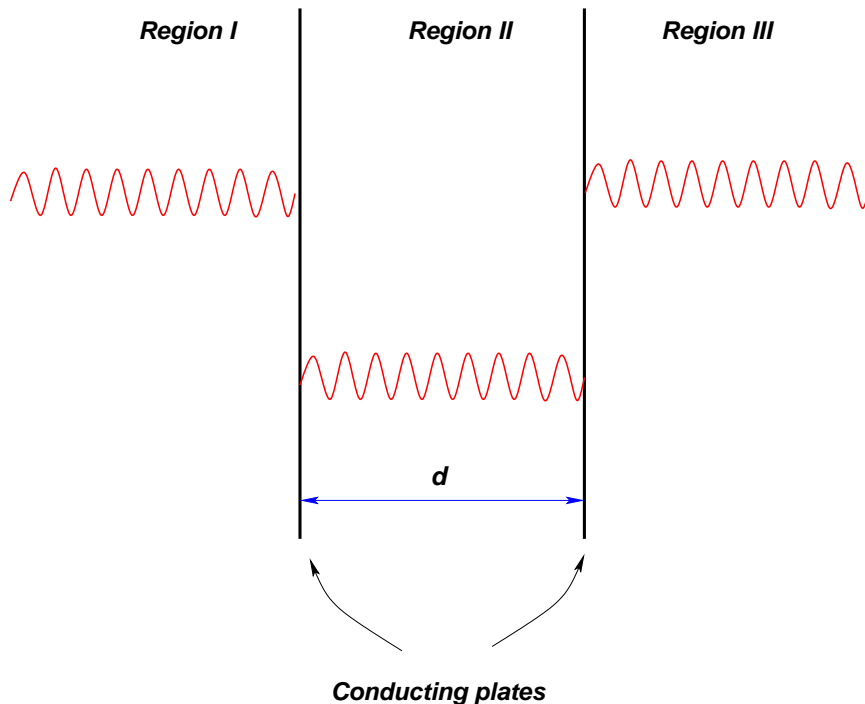
$$:\hat{H}: \equiv \hat{H} - \langle 0 | \hat{H} | 0 \rangle = \int d^3p \omega_p \hat{\alpha}^\dagger(\vec{p}) \hat{\alpha}(\vec{p}) \quad (81)$$

It is interesting to try to make sense of the divergent term in Eq. (80). This term has two sources of divergence. One is associated with the delta function evaluated at zero coming from the fact that we are working in an infinite volume. It can be regularized for large but finite volume by replacing  $\delta(\vec{0}) \sim V$ . Hence, it is of infrared origin. The second one comes from the integration of  $\omega_p$  at large values of the momentum and it is then an ultraviolet divergence. The infrared divergence can be regularized by considering the scalar field to be living in a box of finite volume  $V$ . In this case the vacuum energy is

$$E_{\text{vac}} \equiv \langle 0 | \hat{H} | 0 \rangle = \sum_{\vec{p}} \frac{1}{2} \omega_p. \quad (82)$$

Written in this way the interpretation of the vacuum energy is straightforward. A free scalar quantum field can be seen as an infinite collection of harmonic oscillators per unit volume, each one labelled by  $\vec{p}$ . Even if those oscillators are not excited, they contribute to the vacuum energy with their zero-point energy, given by  $\frac{1}{2}\omega_p$ . This vacuum contribution to the energy adds up to infinity even if we work at finite volume, since even then there are modes with arbitrary high momentum contributing to the sum,  $p_i = \frac{n_i \pi}{L_i}$ , with  $L_i$  the sides of the box of volume  $V$  and  $n_i$  an integer. Hence, this divergence is of ultraviolet origin.

Our discussion leads us to the conclusion that the vacuum in quantum field theory is radically different from the classical idea of the vacuum as “empty space”. Indeed, we have seen that a quantum field can be regarded as a set of an infinite number of harmonic oscillators and that the ground state of the system is obtained when *all* oscillators are in their respective ground states. This being so, we know from elementary quantum mechanics that a harmonic oscillator in its ground state is not “at rest”, but



**Fig. 6:** Illustration of the Casimir effect. In regions I and II the spectrum of modes of the momentum  $p_{\perp}$  is continuous, while in the space between the plates (region II) it is quantized in units of  $\frac{\pi}{d}$ .

fluctuate with an energy given by its zero-point energy. When translated to quantum field theory, this means that the vacuum can be picture as a medium where virtual particles are continuously created and annihilated. As we will see, this nontrivial character of the vacuum has physical consequences ranging from the Casimir effect (see below) to the screening or antiscreening of charges in gauge theories (see Section 8.2).

### 3.2 The Casimir effect

The presence of a vacuum energy is not characteristic of the scalar field. It is also present in other cases, in particular in quantum electrodynamics. Although one might be tempted to discarding this infinite contribution to the energy of the vacuum as unphysical, it has observable consequences. In 1948 Hendrik Casimir pointed out [16] that although a formally divergent vacuum energy would not be observable, any variation in this energy would be (see [17] for comprehensive reviews).

To show this he devised the following experiment. Consider a couple of infinite, perfectly conducting plates placed parallel to each other at a distance  $d$  (see Fig. 6). Because the conducting plates fix the boundary condition of the vacuum modes of the electromagnetic field these are discrete in between the plates (region II), while outside there is a continuous spectrum of modes (regions I and III). In order to calculate the force between the plates we can take the vacuum energy of the electromagnetic field as given by the contribution of two scalar fields corresponding to the two polarizations of the photon. Therefore we can use the formulas derived above.

A naive calculation of the vacuum energy in this system gives a divergent result. This infinity can be removed, however, by subtracting the vacuum energy corresponding to the situation where the plates are removed

$$E(d)_{\text{reg}} = E(d)_{\text{vac}} - E(\infty)_{\text{vac}} \quad (83)$$

This subtraction cancels the contribution of the modes outside the plates. Because of the boundary

conditions imposed by the plates the momentum of the modes perpendicular to the plates are quantized according to  $p_{\perp} = \frac{n\pi}{d}$ , with  $n$  a non-negative integer. If we consider that the size of the plates is much larger than their separation  $d$  we can take the momenta parallel to the plates  $\vec{p}_{\parallel}$  as continuous. For  $n > 0$  we have two polarizations for each vacuum mode of the electromagnetic field, each contributing like  $\frac{1}{2}\sqrt{\vec{p}_{\parallel}^2 + p_{\perp}^2}$  to the vacuum energy. On the other hand, when  $p_{\perp} = 0$  the corresponding modes of the field are effectively (2+1)-dimensional and therefore there is only one polarization. Keeping this in mind, we can write

$$E(d)_{\text{reg}} = S \int \frac{d^2 p_{\parallel}}{(2\pi)^2} \frac{1}{2} |\vec{p}_{\parallel}| + 2S \int \frac{d^2 p_{\parallel}}{(2\pi)^2} \sum_{n=1}^{\infty} \frac{1}{2} \sqrt{\vec{p}_{\parallel}^2 + \left(\frac{n\pi}{d}\right)^2} - 2Sd \int \frac{d^3 p}{(2\pi)^3} \frac{1}{2} |\vec{p}| \quad (84)$$

where  $S$  is the area of the plates. The factors of 2 take into account the two propagating degrees of freedom of the electromagnetic field, as discussed above. In order to ensure the convergence of integrals and infinite sums we can introduce an exponential damping factor<sup>4</sup>

$$E(d)_{\text{reg}} = \frac{1}{2}S \int \frac{d^2 p_{\perp}}{(2\pi)^2} e^{-\frac{1}{\Lambda} |\vec{p}_{\perp}|} |\vec{p}_{\perp}| + S \sum_{n=1}^{\infty} \int \frac{d^2 p_{\parallel}}{(2\pi)^2} e^{-\frac{1}{\Lambda} \sqrt{\vec{p}_{\parallel}^2 + \left(\frac{n\pi}{d}\right)^2}} \sqrt{\vec{p}_{\parallel}^2 + \left(\frac{n\pi}{d}\right)^2} - Sd \int_{-\infty}^{\infty} \frac{dp_{\perp}}{2\pi} \int \frac{d^2 p_{\parallel}}{(2\pi)^2} e^{-\frac{1}{\Lambda} \sqrt{\vec{p}_{\parallel}^2 + p_{\perp}^2}} \sqrt{\vec{p}_{\parallel}^2 + p_{\perp}^2} \quad (85)$$

where  $\Lambda$  is an ultraviolet cutoff. It is now straightforward to see that if we define the function

$$F(x) = \frac{1}{2\pi} \int_0^{\infty} y dy e^{-\frac{1}{\Lambda} \sqrt{y^2 + \left(\frac{x\pi}{d}\right)^2}} \sqrt{y^2 + \left(\frac{x\pi}{d}\right)^2} = \frac{1}{4\pi} \int_{\left(\frac{x\pi}{d}\right)^2}^{\infty} dz e^{-\frac{\sqrt{z}}{\Lambda}} \sqrt{z} \quad (86)$$

the regularized vacuum energy can be written as

$$E(d)_{\text{reg}} = S \left[ \frac{1}{2} F(0) + \sum_{n=1}^{\infty} F(n) - \int_0^{\infty} dx F(x) \right] \quad (87)$$

This expression can be evaluated using the Euler-MacLaurin formula [19]

$$\sum_{n=1}^{\infty} F(n) - \int_0^{\infty} dx F(x) = -\frac{1}{2} [F(0) + F(\infty)] + \frac{1}{12} [F'(\infty) - F'(0)] - \frac{1}{720} [F'''(\infty) - F'''(0)] + \dots \quad (88)$$

Since for our function  $F(\infty) = F'(\infty) = F'''(\infty) = 0$  and  $F'(0) = 0$ , the value of  $E(d)_{\text{reg}}$  is determined by  $F'''(0)$ . Computing this term and removing the ultraviolet cutoff,  $\Lambda \rightarrow \infty$  we find the result

$$E(d)_{\text{reg}} = \frac{S}{720} F'''(0) = -\frac{\pi^2 S}{720 d^3}. \quad (89)$$

Then, the force per unit area between the plates is given by

$$P_{\text{Casimir}} = -\frac{\pi^2}{240} \frac{1}{d^4}. \quad (90)$$

The minus sign shows that the force between the plates is attractive. This is the so-called Casimir effect. It was experimentally measured in 1958 by Sparnaay [18] and since then the Casimir effect has been checked with better and better precision in a variety of situations [17].

<sup>4</sup>Actually, one could introduce any cutoff function  $f(p_{\perp}^2 + p_{\parallel}^2)$  going to zero fast enough as  $p_{\perp}, p_{\parallel} \rightarrow \infty$ . The result is independent of the particular function used in the calculation.

## 4 Theories and Lagrangians

Up to this point we have used a scalar field to illustrate our discussion of the quantization procedure. However, nature is richer than that and it is necessary to consider other fields with more complicated behavior under Lorentz transformations. Before considering other fields we pause and study the properties of the Lorentz group.

### 4.1 Representations of the Lorentz group

In four dimensions the Lorentz group has six generators. Three of them correspond to the generators of the group of rotations in three dimensions  $SO(3)$ . In terms of the generators  $J_i$  of the group a finite rotation of angle  $\varphi$  with respect to an axis determined by a unitary vector  $\vec{e}$  can be written as

$$R(\vec{e}, \varphi) = e^{-i\varphi \vec{e} \cdot \vec{J}}, \quad \vec{J} = \begin{pmatrix} J_1 \\ J_2 \\ J_3 \end{pmatrix}. \quad (91)$$

The other three generators of the Lorentz group are associated with boosts  $M_i$  along the three spatial directions. A boost with rapidity  $\lambda$  along a direction  $\vec{u}$  is given by

$$B(\vec{u}, \lambda) = e^{-i\lambda \vec{u} \cdot \vec{M}}, \quad \vec{M} = \begin{pmatrix} M_1 \\ M_2 \\ M_3 \end{pmatrix}. \quad (92)$$

These six generators satisfy the algebra

$$\begin{aligned} [J_i, J_j] &= i\epsilon_{ijk} J_k, \\ [J_i, M_k] &= i\epsilon_{ijk} M_k, \\ [M_i, M_j] &= -i\epsilon_{ijk} J_k, \end{aligned} \quad (93)$$

The first line corresponds to the commutation relations of  $SO(3)$ , while the second one implies that the generators of the boosts transform like a vector under rotations.

At first sight, to find representations of the algebra (93) might seem difficult. The problem is greatly simplified if we consider the following combination of the generators

$$J_k^\pm = \frac{1}{2}(J_k \pm iM_k). \quad (94)$$

Using (93) it is easy to prove that the new generators  $J_k^\pm$  satisfy the algebra

$$\begin{aligned} [J_i^\pm, J_j^\pm] &= i\epsilon_{ijk} J_k^\pm, \\ [J_i^+, J_j^-] &= 0. \end{aligned} \quad (95)$$

Then the Lorentz algebra (93) is actually equivalent to two copies of the algebra of  $SU(2) \approx SO(3)$ . Therefore the irreducible representations of the Lorentz group can be obtained from the well-known representations of  $SU(2)$ . Since the latter ones are labelled by the spin  $s = k + \frac{1}{2}, k$  (with  $k \in \mathbb{N}$ ), any representation of the Lorentz algebra can be identified by specifying  $(s_+, s_-)$ , the spins of the representations of the two copies of  $SU(2)$  that made up the algebra (93).

To get familiar with this way of labelling the representations of the Lorentz group we study some particular examples. Let us start with the simplest one  $(s_+, s_-) = (0, 0)$ . This state is a singlet under  $J_i^\pm$  and therefore also under rotations and boosts. Therefore we have a scalar.

The next interesting cases are  $(\frac{1}{2}, 0)$  and  $(0, \frac{1}{2})$ . They correspond respectively to a right-handed and a left-handed Weyl spinor. Their properties will be studied in more detail below. In the case of

Representation	Type of field
$(\mathbf{0}, \mathbf{0})$	Scalar
$(\frac{1}{2}, \mathbf{0})$	Right-handed spinor
$(\mathbf{0}, \frac{1}{2})$	Left-handed spinor
$(\frac{1}{2}, \frac{1}{2})$	Vector
$(\mathbf{1}, \mathbf{0})$	Selfdual antisymmetric 2-tensor
$(\mathbf{0}, \mathbf{1})$	Anti-selfdual antisymmetric 2-tensor

**Table 1:** Representations of the Lorentz group

$(\frac{1}{2}, \frac{1}{2})$ , since from Eq. (94) we see that  $J_i = J_i^+ + J_i^-$  the rules of addition of angular momentum tell us that there are two states, one of them transforming as a vector and another one as a scalar under three-dimensional rotations. Actually, a more detailed analysis shows that the singlet state corresponds to the time component of a vector and the states combine to form a vector under the Lorentz group.

There are also more “exotic” representations. For example we can consider the  $(\mathbf{1}, \mathbf{0})$  and  $(\mathbf{0}, \mathbf{1})$  representations corresponding respectively to a selfdual and an anti-selfdual rank-two antisymmetric tensor. In Table 1 we summarize the previous discussion.

To conclude our discussion of the representations of the Lorentz group we notice that under a parity transformation the generators of  $SO(1,3)$  transform as

$$P : J_i \longrightarrow J_i, \quad P : M_i \longrightarrow -M_i \quad (96)$$

this means that  $P : J_i^\pm \longrightarrow J_i^\mp$  and therefore a representation  $(\mathbf{s}_1, \mathbf{s}_2)$  is transformed into  $(\mathbf{s}_2, \mathbf{s}_1)$ . This means that, for example, a vector  $(\frac{1}{2}, \frac{1}{2})$  is invariant under parity, whereas a left-handed Weyl spinor  $(\frac{1}{2}, \mathbf{0})$  transforms into a right-handed one  $(\mathbf{0}, \frac{1}{2})$  and vice versa.

## 4.2 Spinors

**Weyl spinors.** Let us go back to the two spinor representations of the Lorentz group, namely  $(\frac{1}{2}, \mathbf{0})$  and  $(\mathbf{0}, \frac{1}{2})$ . These representations can be explicitly constructed using the Pauli matrices as

$$\begin{aligned} J_i^+ &= \frac{1}{2}\sigma^i, & J_i^- &= 0 & \text{for } & (\frac{1}{2}, \mathbf{0}), \\ J_i^+ &= 0, & J_i^- &= \frac{1}{2}\sigma^i & \text{for } & (\mathbf{0}, \frac{1}{2}). \end{aligned} \quad (97)$$

We denote by  $u_\pm$  a complex two-component object that transforms in the representation  $\mathbf{s}_\pm = \frac{1}{2}$  of  $J_\pm^i$ . If we define  $\sigma_\pm^\mu = (\mathbf{1}, \pm\sigma^i)$  we can construct the following vector quantities

$$u_+^\dagger \sigma_+^\mu u_+, \quad u_-^\dagger \sigma_-^\mu u_-. \quad (98)$$

Notice that since  $(J_i^\pm)^\dagger = J_i^\mp$  the hermitian conjugated fields  $u_\pm^\dagger$  are in the  $(\mathbf{0}, \frac{1}{2})$  and  $(\frac{1}{2}, \mathbf{0})$  respectively.

To construct a free Lagrangian for the fields  $u_\pm$  we have to look for quadratic combinations of the fields that are Lorentz scalars. If we also demand invariance under global phase rotations

$$u_\pm \longrightarrow e^{i\theta} u_\pm \quad (99)$$

we are left with just one possibility up to a sign

$$\mathcal{L}_{\text{Weyl}}^{\pm} = iu_{\pm}^{\dagger} \left( \partial_t \pm \vec{\sigma} \cdot \vec{\nabla} \right) u_{\pm} = iu_{\pm}^{\dagger} \sigma_{\pm}^{\mu} \partial_{\mu} u_{\pm}. \quad (100)$$

This is the Weyl Lagrangian. In order to grasp the physical meaning of the spinors  $u_{\pm}$  we write the equations of motion

$$\left( \partial_0 \pm \vec{\sigma} \cdot \vec{\nabla} \right) u_{\pm} = 0. \quad (101)$$

Multiplying this equation on the left by  $\left( \partial_0 \mp \vec{\sigma} \cdot \vec{\nabla} \right)$  and applying the algebraic properties of the Pauli matrices we conclude that  $u_{\pm}$  satisfies the massless Klein-Gordon equation

$$\partial_{\mu} \partial^{\mu} u_{\pm} = 0, \quad (102)$$

whose solutions are:

$$u_{\pm}(x) = u_{\pm}(k) e^{-ik \cdot x}, \quad \text{with } k^0 = |\vec{k}|. \quad (103)$$

Plugging these solutions back into the equations of motion (101) we find

$$\left( |\vec{k}| \mp \vec{k} \cdot \vec{\sigma} \right) u_{\pm} = 0, \quad (104)$$

which implies

$$\begin{aligned} u_{+} : \quad & \frac{\vec{\sigma} \cdot \vec{k}}{|\vec{k}|} = 1, \\ u_{-} : \quad & \frac{\vec{\sigma} \cdot \vec{k}}{|\vec{k}|} = -1. \end{aligned} \quad (105)$$

Since the spin operator is defined as  $\vec{s} = \frac{1}{2} \vec{\sigma}$ , the previous expressions give the chirality of the states with wave function  $u_{\pm}$ , i.e. the projection of spin along the momentum of the particle. Therefore we conclude that  $u_{+}$  is a Weyl spinor of positive helicity  $\lambda = \frac{1}{2}$ , while  $u_{-}$  has negative helicity  $\lambda = -\frac{1}{2}$ . This agrees with our assertion that the representation  $(\frac{1}{2}, \mathbf{0})$  corresponds to a right-handed Weyl fermion (positive chirality) whereas  $(\mathbf{0}, \frac{1}{2})$  is a left-handed Weyl fermion (negative chirality). For example, in the standard model neutrinos are left-handed Weyl spinors and therefore transform in the representation  $(\mathbf{0}, \frac{1}{2})$  of the Lorentz group.

Nevertheless, it is possible that we were too restrictive in constructing the Weyl Lagrangian (100). There we constructed the invariants from the vector bilinears (98) corresponding to the product representations

$$\left( \frac{1}{2}, \frac{1}{2} \right) = \left( \frac{1}{2}, \mathbf{0} \right) \otimes \left( \mathbf{0}, \frac{1}{2} \right) \quad \text{and} \quad \left( \frac{1}{2}, \frac{1}{2} \right) = \left( \mathbf{0}, \frac{1}{2} \right) \otimes \left( \frac{1}{2}, \mathbf{0} \right). \quad (106)$$

In particular our insistence in demanding the Lagrangian to be invariant under the global symmetry  $u_{\pm} \rightarrow e^{i\theta} u_{\pm}$  rules out the scalar term that appears in the product representations

$$\left( \frac{1}{2}, \mathbf{0} \right) \otimes \left( \frac{1}{2}, \mathbf{0} \right) = \left( \mathbf{1}, \mathbf{0} \right) \oplus \left( \mathbf{0}, \mathbf{0} \right), \quad \left( \mathbf{0}, \frac{1}{2} \right) \otimes \left( \mathbf{0}, \frac{1}{2} \right) = \left( \mathbf{0}, \mathbf{1} \right) \oplus \left( \mathbf{0}, \mathbf{0} \right). \quad (107)$$

The singlet representations corresponds to the antisymmetric combinations

$$\epsilon_{ab} u_{\pm}^a u_{\pm}^b, \quad (108)$$

where  $\epsilon_{ab}$  is the antisymmetric symbol  $\epsilon_{12} = -\epsilon_{21} = 1$ .

At first sight it might seem that the term (108) vanishes identically because of the antisymmetry of the  $\epsilon$ -symbol. However we should keep in mind that the spin-statistic theorem (more on this later) demands that fields with half-integer spin have to satisfy the Fermi-Dirac statistics and therefore satisfy anticommutation relations, whereas fields of integer spin follow the statistic of Bose-Einstein and, as a consequence, quantization replaces Poisson brackets by commutators. This implies that the components of the Weyl fermions  $u_{\pm}$  are anticommuting Grassmann fields

$$u_{\pm}^a u_{\pm}^b + u_{\pm}^b u_{\pm}^a = 0. \quad (109)$$

It is important to realize that, strictly speaking, fermions (i.e., objects that satisfy the Fermi-Dirac statistics) do not exist classically. The reason is that they satisfy the Pauli exclusion principle and therefore each quantum state can be occupied, at most, by one fermion. Therefore the naïve definition of the classical limit as a limit of large occupation numbers cannot be applied. Fermion field do not really make sense classically.

Since the combination (108) does not vanish and we can construct a new Lagrangian

$$\mathcal{L}_{\text{Weyl}}^{\pm} = iu_{\pm}^{\dagger} \sigma_{\pm}^{\mu} \partial_{\mu} u_{\pm} - \frac{m}{2} \epsilon_{ab} u_{\pm}^a u_{\pm}^b + \text{h.c.} \quad (110)$$

This mass term, called of Majorana type, is allowed if we do not worry about breaking the global U(1) symmetry  $u_{\pm} \rightarrow e^{i\theta} u_{\pm}$ . This is not the case, for example, of charged chiral fermions, since the Majorana mass violates the conservation of electric charge or any other gauge U(1) charge. In the standard model, however, there is no such a problem if we introduce Majorana masses for right-handed neutrinos, since they are singlet under all standard model gauge groups. Such a term will break, however, the global U(1) lepton number charge because the operator  $\epsilon_{ab} \nu_R^a \nu_R^b$  changes the lepton number by two units

**Dirac spinors.** We have seen that parity interchanges the representations  $(\frac{1}{2}, \mathbf{0})$  and  $(\mathbf{0}, \frac{1}{2})$ , i.e. it changes right-handed with left-handed fermions

$$P : u_{\pm} \longrightarrow u_{\mp}. \quad (111)$$

An obvious way to build a parity invariant theory is to introduce a pair of Weyl fermions  $u_{+}$  and  $u_{-}$ . Actually, these two fields can be combined in a single four-component spinor

$$\psi = \begin{pmatrix} u_{+} \\ u_{-} \end{pmatrix} \quad (112)$$

transforming in the reducible representation  $(\frac{1}{2}, \mathbf{0}) \oplus (\mathbf{0}, \frac{1}{2})$ .

Since now we have both  $u_{+}$  and  $u_{-}$  simultaneously at our disposal the equations of motion for  $u_{\pm}$ ,  $i\sigma_{\pm}^{\mu} \partial_{\mu} u_{\pm} = 0$  can be modified, while keeping them linear, to

$$\left. \begin{array}{l} i\sigma_{+}^{\mu} \partial_{\mu} u_{+} = m u_{-} \\ i\sigma_{-}^{\mu} \partial_{\mu} u_{-} = m u_{+} \end{array} \right\} \implies i \begin{pmatrix} \sigma_{+}^{\mu} & 0 \\ 0 & \sigma_{-}^{\mu} \end{pmatrix} \partial_{\mu} \psi = m \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix} \psi. \quad (113)$$

These equations of motion can be derived from the Lagrangian density

$$\mathcal{L}_{\text{Dirac}} = i\psi^{\dagger} \begin{pmatrix} \sigma_{+}^{\mu} & 0 \\ 0 & \sigma_{-}^{\mu} \end{pmatrix} \partial_{\mu} \psi - m\psi^{\dagger} \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix} \psi. \quad (114)$$

To simplify the notation it is useful to define the Dirac  $\gamma$ -matrices as

$$\gamma^{\mu} = \begin{pmatrix} 0 & \sigma_{-}^{\mu} \\ \sigma_{+}^{\mu} & 0 \end{pmatrix} \quad (115)$$



and the Dirac conjugate spinor  $\bar{\psi}$

$$\bar{\psi} \equiv \psi^\dagger \gamma^0 = \psi^\dagger \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}. \quad (116)$$

Now the Lagrangian (114) can be written in the more compact form

$$\mathcal{L}_{\text{Dirac}} = \bar{\psi} (i\gamma^\mu \partial_\mu - m) \psi. \quad (117)$$

The associated equations of motion give the Dirac equation (11) with the identifications

$$\gamma^0 = \beta, \quad \gamma^i = i\alpha^i. \quad (118)$$

In addition, the  $\gamma$ -matrices defined in (115) satisfy the Clifford algebra

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}. \quad (119)$$

In  $D$  dimensions this algebra admits representations of dimension  $2^{\lfloor \frac{D}{2} \rfloor}$ . When  $D$  is even the Dirac fermions  $\psi$  transform in a reducible representation of the Lorentz group. In the case of interest,  $D = 4$  this is easy to prove by defining the matrix

$$\gamma^5 = -i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} \mathbf{1} & 0 \\ 0 & -\mathbf{1} \end{pmatrix}. \quad (120)$$

We see that  $\gamma^5$  anticommutes with all other  $\gamma$ -matrices. This implies that

$$[\gamma^5, \sigma^{\mu\nu}] = 0, \quad \text{with} \quad \sigma^{\mu\nu} = -\frac{i}{4}[\gamma^\mu, \gamma^\nu]. \quad (121)$$

Because of Schur's lemma (see Appendix) this implies that the representation of the Lorentz group provided by  $\sigma^{\mu\nu}$  is reducible into subspaces spanned by the eigenvectors of  $\gamma^5$  with the same eigenvalue. If we define the projectors  $P_\pm = \frac{1}{2}(1 \pm \gamma^5)$  these subspaces correspond to

$$P_+\psi = \begin{pmatrix} u_+ \\ 0 \end{pmatrix}, \quad P_-\psi = \begin{pmatrix} 0 \\ u_- \end{pmatrix}, \quad (122)$$

which are precisely the Weyl spinors introduced before.

Our next task is to quantize the Dirac Lagrangian. This will be done along the lines used for the Klein-Gordon field, starting with a general solution to the Dirac equation and introducing the corresponding set of creation-annihilation operators. Therefore we start by looking for a complete basis of solutions to the Dirac equation. In the case of the scalar field the elements of the basis were labelled by their four-momentum  $k^\mu$ . Now, however, we have more degrees of freedom since we are dealing with a spinor which means that we have to add extra labels. Looking back at Eq. (105) we can define the helicity operator for a Dirac spinor as

$$\lambda = \frac{1}{2} \vec{\sigma} \cdot \frac{\vec{k}}{|\vec{k}|} \begin{pmatrix} \mathbf{1} & 0 \\ 0 & \mathbf{1} \end{pmatrix}. \quad (123)$$

Hence, each element of the basis of functions is labelled by its four-momentum  $k^\mu$  and the corresponding eigenvalue  $s$  of the helicity operator. For positive energy solutions we then propose the ansatz

$$u(k, s)e^{-ik \cdot x}, \quad s = \pm \frac{1}{2}, \quad (124)$$

where  $u_\alpha(k, s)$  ( $\alpha = 1, \dots, 4$ ) is a four-component spinor. Substituting in the Dirac equation we obtain

$$(\not{k} - m)u(k, s) = 0. \quad (125)$$

In the same way, for negative energy solutions we have

$$v(k, s)e^{ik \cdot x}, \quad s = \pm \frac{1}{2}, \quad (126)$$

where  $v(k, s)$  has to satisfy

$$(\not{k} + m)v(k, s) = 0. \quad (127)$$

Multiplying Eqs. (125) and (127) on the left respectively by  $(\not{k} \mp m)$  we find that the momentum is on the mass shell,  $k^2 = m^2$ . Because of this, the wave function for both positive- and negative-energy solutions can be labeled as well using the three-momentum  $\vec{k}$  of the particle,  $u(\vec{k}, s)$ ,  $v(\vec{k}, s)$ .

A detailed analysis shows that the functions  $u(\vec{k}, s)$ ,  $v(\vec{k}, s)$  satisfy the properties

$$\begin{aligned} \bar{u}(\vec{k}, s)u(\vec{k}, s) &= 2m, & \bar{v}(\vec{k}, s)v(\vec{k}, s) &= -2m, \\ \bar{u}(\vec{k}, s)\gamma^\mu u(\vec{k}, s) &= 2k^\mu, & \bar{v}(\vec{k}, s)\gamma^\mu v(\vec{k}, s) &= 2k^\mu, \\ \sum_{s=\pm\frac{1}{2}} u_\alpha(\vec{k}, s)\bar{u}_\beta(\vec{k}, s) &= (\not{k} + m)_{\alpha\beta}, & \sum_{s=\pm\frac{1}{2}} v_\alpha(\vec{k}, s)\bar{v}_\beta(\vec{k}, s) &= (\not{k} - m)_{\alpha\beta}, \end{aligned} \quad (128)$$

with  $k^0 = \omega_k = \sqrt{\vec{k}^2 + m^2}$ . Then, a general solution to the Dirac equation including creation and annihilation operators can be written as:

$$\hat{\psi}(t, \vec{x}) = \int \frac{d^3k}{(2\pi)^3} \frac{1}{2\omega_k} \sum_{s=\pm\frac{1}{2}} \left[ u(\vec{k}, s)\hat{b}(\vec{k}, s)e^{-i\omega_k t + i\vec{k} \cdot \vec{x}} + v(\vec{k}, s)\hat{d}^\dagger(\vec{k}, s)e^{i\omega_k t - i\vec{k} \cdot \vec{x}} \right]. \quad (129)$$

The operators  $\hat{b}^\dagger(\vec{k}, s)$ ,  $\hat{b}(\vec{k}, s)$  respectively create and annihilate a spin- $\frac{1}{2}$  particle (for example, an electron) out of the vacuum with momentum  $\vec{k}$  and helicity  $s$ . Because we are dealing with half-integer spin fields, the spin-statistics theorem forces canonical anticommutation relations for  $\hat{\psi}$  which means that the creation-annihilation operators satisfy the algebra<sup>5</sup>

$$\begin{aligned} \{b(\vec{k}, s), b^\dagger(\vec{k}', s')\} &= \delta(\vec{k} - \vec{k}')\delta_{ss'}, \\ \{b(\vec{k}, s), b(\vec{k}', s')\} &= \{b^\dagger(\vec{k}, s), b^\dagger(\vec{k}', s')\} = 0. \end{aligned} \quad (130)$$

In the case of  $d(\vec{k}, s)$ ,  $d^\dagger(\vec{k}, s)$  we have a set of creation-annihilation operators for the corresponding antiparticles (for example positrons). This is clear if we notice that  $d^\dagger(\vec{k}, s)$  can be seen as the annihilation operator of a negative energy state of the Dirac equation with wave function  $v_\alpha(\vec{k}, s)$ . As we saw, in the Dirac sea picture this corresponds to the creation of an antiparticle out of the vacuum (see Fig. 2). The creation-annihilation operators for antiparticles also satisfy the fermionic algebra

$$\begin{aligned} \{d(\vec{k}, s), d^\dagger(\vec{k}', s')\} &= \delta(\vec{k} - \vec{k}')\delta_{ss'}, \\ \{d(\vec{k}, s), d(\vec{k}', s')\} &= \{d^\dagger(\vec{k}, s), d^\dagger(\vec{k}', s')\} = 0. \end{aligned} \quad (131)$$

All other anticommutators between  $b(\vec{k}, s)$ ,  $b^\dagger(\vec{k}, s)$  and  $d(\vec{k}, s)$ ,  $d^\dagger(\vec{k}, s)$  vanish.

The Hamiltonian operator for the Dirac field is

$$\hat{H} = \frac{1}{2} \sum_{s=\pm\frac{1}{2}} \int \frac{d^3k}{(2\pi)^3} \left[ b^\dagger(\vec{k}, s)b(\vec{k}, s) - d(\vec{k}, s)d^\dagger(\vec{k}, s) \right]. \quad (132)$$

At this point we realize again of the necessity of quantizing the theory using anticommutators instead of commutators. Had we use canonical commutation relations, the second term inside the integral in

<sup>5</sup>To simplify notation, and since there is no risk of confusion, we drop from now on the hat to indicate operators.

(132) would give the number operator  $d^\dagger(\vec{k}, s)d(\vec{k}, s)$  with a minus sign in front. As a consequence the Hamiltonian would be unbounded from below and we would be facing again the instability of the theory already noticed in the context of relativistic quantum mechanics. However, because of the *anticommutation* relations (131), the Hamiltonian (132) takes the form

$$\hat{H} = \sum_{s=\pm\frac{1}{2}} \int \frac{d^3k}{(2\pi)^3} \frac{1}{2\omega_k} \left[ \omega_k b^\dagger(\vec{k}, s)b(\vec{k}, s) + \omega_k d^\dagger(\vec{k}, s)d(\vec{k}, s) \right] - 2 \int d^3k \omega_k \delta(\vec{0}). \quad (133)$$

As with the scalar field, we find a divergent vacuum energy contribution due to the zero-point energy of the infinite number of harmonic oscillators. Unlike the Klein-Gordon field, the vacuum energy is negative. In section 9.2 we will see that in certain type of theories called supersymmetric, where the number of bosonic and fermionic degrees of freedom is the same, there is a cancellation of the vacuum energy. The divergent contribution can be removed by the normal order prescription

$$:\hat{H}: = \sum_{s=\pm\frac{1}{2}} \int \frac{d^3k}{(2\pi)^3} \frac{1}{2\omega_k} \left[ \omega_k b^\dagger(\vec{k}, s)b(\vec{k}, s) + \omega_k d^\dagger(\vec{k}, s)d(\vec{k}, s) \right]. \quad (134)$$

Finally, let us mention that using the Dirac equation it is easy to prove that there is a conserved four-current given by

$$j^\mu = \bar{\psi}\gamma^\mu\psi, \quad \partial_\mu j^\mu = 0. \quad (135)$$

As we will explain further in sec. 6 this current is associated to the invariance of the Dirac Lagrangian under the global phase shift  $\psi \rightarrow e^{i\theta}\psi$ . In electrodynamics the associated conserved charge

$$Q = e \int d^3x j^0 \quad (136)$$

is identified with the electric charge.

### 4.3 Gauge fields

In classical electrodynamics the basic quantities are the electric and magnetic fields  $\vec{E}, \vec{B}$ . These can be expressed in terms of the scalar and vector potential  $(\varphi, \vec{A})$

$$\begin{aligned} \vec{E} &= -\vec{\nabla}\varphi - \frac{\partial\vec{A}}{\partial t}, \\ \vec{B} &= \vec{\nabla} \times \vec{A}. \end{aligned} \quad (137)$$

From these equations it follows that there is an ambiguity in the definition of the potentials given by the gauge transformations

$$\varphi(t, \vec{x}) \rightarrow \varphi(t, \vec{x}) + \frac{\partial}{\partial t}\epsilon(t, \vec{x}), \quad \vec{A}(t, \vec{x}) \rightarrow \vec{A}(t, \vec{x}) - \vec{\nabla}\epsilon(t, \vec{x}). \quad (138)$$

Classically  $(\varphi, \vec{A})$  are seen as only a convenient way to solve the Maxwell equations, but without physical relevance.

The equations of electrodynamics can be recast in a manifestly Lorentz invariant form using the four-vector gauge potential  $A^\mu = (\varphi, \vec{A})$  and the antisymmetric rank-two tensor:  $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ . Maxwell's equations become

$$\begin{aligned} \partial_\mu F^{\mu\nu} &= j^\nu, \\ \epsilon^{\mu\nu\sigma\eta} \partial_\nu F_{\sigma\eta} &= 0, \end{aligned} \quad (139)$$

where the four-current  $j^\mu = (\rho, \vec{j})$  contains the charge density and the electric current. The field strength tensor  $F_{\mu\nu}$  and the Maxwell equations are invariant under gauge transformations (138), which in covariant form read

$$A_\mu \longrightarrow A_\mu + \partial_\mu \epsilon. \quad (140)$$

Finally, the equations of motion of charged particles are given, in covariant form, by

$$m \frac{du^\mu}{d\tau} = e F^{\mu\nu} u_\nu, \quad (141)$$

where  $e$  is the charge of the particle and  $u^\mu(\tau)$  its four-velocity as a function of the proper time.

The physical rôle of the vector potential becomes manifest only in Quantum Mechanics. Using the prescription of minimal substitution  $\vec{p} \rightarrow \vec{p} - e\vec{A}$ , the Schrödinger equation describing a particle with charge  $e$  moving in an electromagnetic field is

$$i\partial_t \Psi = \left[ -\frac{1}{2m} \left( \vec{\nabla} - ie\vec{A} \right)^2 + e\varphi \right] \Psi. \quad (142)$$

Because of the explicit dependence on the electromagnetic potentials  $\varphi$  and  $\vec{A}$ , this equation seems to change under the gauge transformations (138). This is physically acceptable only if the ambiguity does not affect the probability density given by  $|\Psi(t, \vec{x})|^2$ . Therefore, a gauge transformation of the electromagnetic potential should amount to a change in the (unobservable) phase of the wave function. This is indeed what happens: the Schrödinger equation (142) is invariant under the gauge transformations (138) provided the phase of the wave function is transformed at the same time according to

$$\Psi(t, \vec{x}) \longrightarrow e^{-ie\epsilon(t, \vec{x})} \Psi(t, \vec{x}). \quad (143)$$

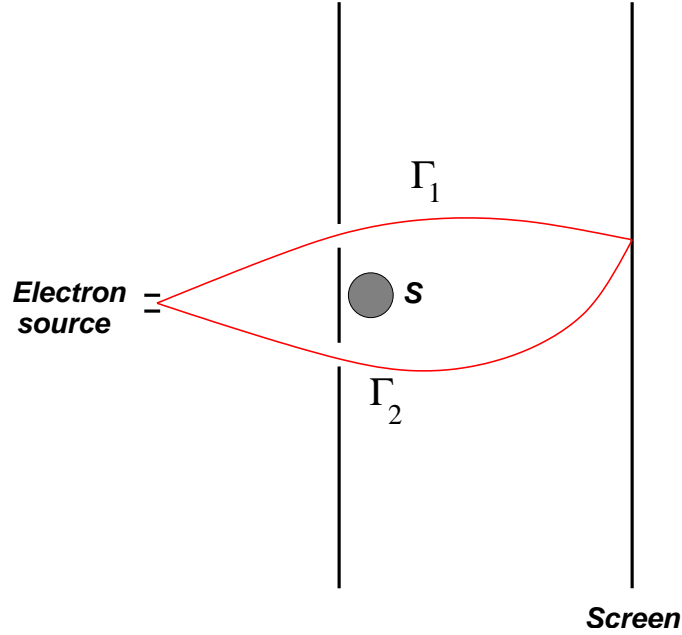
**Aharonov-Bohm effect.** This interplay between gauge transformations and the phase of the wave function give rise to surprising phenomena. The first evidence of the rôle played by the electromagnetic potentials at the quantum level was pointed out by Yakir Aharonov and David Bohm [20]. Let us consider a double slit experiment as shown in Fig. 7, where we have placed a shielded solenoid just behind the first screen. Although the magnetic field is confined to the interior of the solenoid, the vector potential is nonvanishing also outside. Of course the value of  $\vec{A}$  outside the solenoid is a pure gauge, i.e.  $\vec{\nabla} \times \vec{A} = \vec{0}$ , however because the region outside the solenoid is not simply connected the vector potential cannot be gauged to zero everywhere. If we denote by  $\Psi_1^{(0)}$  and  $\Psi_2^{(0)}$  the wave functions for each of the two electron beams in the absence of the solenoid, the total wave function once the magnetic field is switched on can be written as

$$\begin{aligned} \Psi &= e^{ie \int_{\Gamma_1} \vec{A} \cdot d\vec{x}} \Psi_1^{(0)} + e^{ie \int_{\Gamma_2} \vec{A} \cdot d\vec{x}} \Psi_2^{(0)} \\ &= e^{ie \int_{\Gamma_1} \vec{A} \cdot d\vec{x}} \left[ \Psi_1^{(0)} + e^{ie \oint_{\Gamma} \vec{A} \cdot d\vec{x}} \Psi_2^{(0)} \right], \end{aligned} \quad (144)$$

where  $\Gamma_1$  and  $\Gamma_2$  are two curves surrounding the solenoid from different sides, and  $\Gamma$  is any closed loop surrounding it. Therefore the relative phase between the two beams gets an extra term depending on the value of the vector potential outside the solenoid as

$$U = \exp \left[ ie \oint_{\Gamma} \vec{A} \cdot d\vec{x} \right]. \quad (145)$$

Because of the change in the relative phase of the electron wave functions, the presence of the vector potential becomes observable even if the electrons do not feel the magnetic field. If we perform the double-slit experiment when the magnetic field inside the solenoid is switched off we will observe the



**Fig. 7:** Illustration of an interference experiment to show the Aharonov-Bohm effect.  $S$  represent the solenoid in whose interior the magnetic field is confined.

usual interference pattern on the second screen. However if now the magnetic field is switched on, because of the phase (144), a change in the interference pattern will appear. This is the Aharonov-Bohm effect.

The first question that comes up is what happens with gauge invariance. Since we said that  $\vec{A}$  can be changed by a gauge transformation it seems that the resulting interference patterns might depend on the gauge used. Actually, the phase  $U$  in (145) is independent of the gauge although, unlike other gauge-invariant quantities like  $\vec{E}$  and  $\vec{B}$ , is nonlocal. Notice that, since  $\vec{\nabla} \times \vec{A} = \vec{0}$  outside the solenoid, the value of  $U$  does not change under continuous deformations of the closed curve  $\Gamma$ , so long as it does not cross the solenoid.

**The Dirac monopole.** It is very easy to check that the vacuum Maxwell equations remain invariant under the transformation

$$\vec{E} - i\vec{B} \longrightarrow e^{i\theta}(\vec{E} - i\vec{B}), \quad \theta \in [0, 2\pi] \quad (146)$$

which, in particular, for  $\theta = \frac{\pi}{2}$  interchanges the electric and the magnetic fields:  $\vec{E} \rightarrow \vec{B}$ ,  $\vec{B} \rightarrow -\vec{E}$ . This duality symmetry is however broken in the presence of electric sources. Nevertheless the Maxwell equations can be “completed” by introducing sources for the magnetic field  $(\rho_m, \vec{j}_m)$  in such a way that the duality (146) is restored when supplemented by the transformation

$$\rho - i\rho_m \longrightarrow e^{i\theta}(\rho - i\rho_m), \quad \vec{j} - i\vec{j}_m \longrightarrow e^{i\theta}(\vec{j} - i\vec{j}_m). \quad (147)$$

Again for  $\theta = \pi/2$  the electric and magnetic sources get interchanged.

In 1931 Dirac [21] studied the possibility of finding solutions of the completed Maxwell equation with a magnetic monopoles of charge  $g$ , i.e. solutions to

$$\vec{\nabla} \cdot \vec{B} = g \delta(\vec{x}). \quad (148)$$

Away from the position of the monopole  $\vec{\nabla} \cdot \vec{B} = 0$  and the magnetic field can be still derived locally from a vector potential  $\vec{A}$  according to  $\vec{B} = \vec{\nabla} \times \vec{A}$ . However, the vector potential cannot be regular

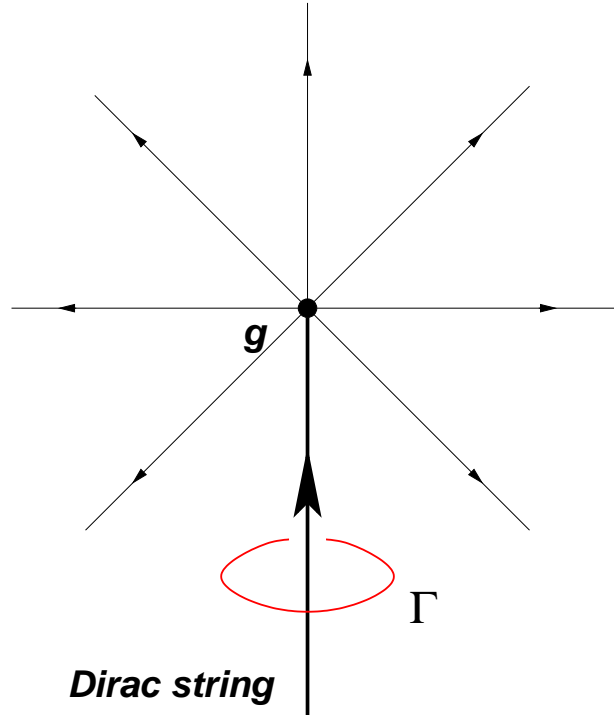


Fig. 8: The Dirac monopole.

everywhere since otherwise Gauss law would imply that the magnetic flux threading a closed surface around the monopole should vanish, contradicting (148).

We look now for solutions to Eq. (148). Working in spherical coordinates we find

$$B_r = \frac{g}{|\vec{x}|^2}, \quad B_\varphi = B_\theta = 0. \quad (149)$$

Away from the position of the monopole ( $\vec{x} \neq \vec{0}$ ) the magnetic field can be derived from the vector potential

$$A_\varphi = \frac{g}{|\vec{x}|} \tan \frac{\theta}{2}, \quad A_r = A_\theta = 0. \quad (150)$$

As expected we find that this vector potential is actually singular around the half-line  $\theta = \pi$  (see Fig. 8). This singular line starting at the position of the monopole is called the Dirac string and its position changes with a change of gauge but cannot be eliminated by any gauge transformation. Physically we can see it as an infinitely thin solenoid confining a magnetic flux entering into the magnetic monopole from infinity that equals the outgoing magnetic flux from the monopole.

Since the position of the Dirac string depends on the gauge chosen it seems that the presence of monopoles introduces an ambiguity. This would be rather strange, since Maxwell equations are gauge invariant also in the presence of magnetic sources. The solution to this apparent riddle lies in the fact that the Dirac string does not pose any consistency problem as far as it does not produce any physical effect, i.e. if its presence turns out to be undetectable. From our discussion of the Aharonov-Bohm effect we know that the wave function of charged particles pick up a phase (145) when surrounding a region where magnetic flux is confined (for example the solenoid in the Aharonov-Bohm experiment). As explained above, the Dirac string associated with the monopole can be seen as a infinitely thin solenoid. Therefore the Dirac string will be unobservable if the phase picked up by the wave function of a charged particle is equal to one. A simple calculation shows that this happens if

$$e^{ie g} = 1 \quad \implies \quad e g = 2\pi n \quad \text{with } n \in \mathbb{Z}. \quad (151)$$

Interestingly, this discussion leads to the conclusion that the presence of a single magnetic monopoles somewhere in the Universe implies for consistency the quantization of the electric charge in units of  $\frac{2\pi}{g}$ , where  $g$  the magnetic charge of the monopole.

**Quantization of the electromagnetic field.** We now proceed to the quantization of the electromagnetic field in the absence of sources  $\rho = 0$ ,  $\vec{j} = \vec{0}$ . In this case the Maxwell equations (139) can be derived from the Lagrangian density

$$\mathcal{L}_{\text{Maxwell}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} = \frac{1}{2}(\vec{E}^2 - \vec{B}^2). \quad (152)$$

Although in general the procedure to quantize the Maxwell Lagrangian is not very different from the one used for the Klein-Gordon or the Dirac field, here we need to deal with a new ingredient: gauge invariance. Unlike the cases studied so far, here the photon field  $A_\mu$  is not unambiguously defined because the action and the equations of motion are insensitive to the gauge transformations  $A_\mu \rightarrow A_\mu + \partial_\mu \varepsilon$ . A first consequence of this symmetry is that the theory has less physical degrees of freedom than one would expect from the fact that we are dealing with a vector field.

The way to tackle the problem of gauge invariance is to fix the freedom in choosing the electromagnetic potential before quantization. This can be done in several ways, for example by imposing the Lorentz gauge fixing condition

$$\partial_\mu A^\mu = 0. \quad (153)$$

Notice that this condition does not fix completely the gauge freedom since Eq. (153) is left invariant by gauge transformations satisfying  $\partial_\mu \partial^\mu \varepsilon = 0$ . One of the advantages, however, of the Lorentz gauge is that it is covariant and therefore does not pose any danger to the Lorentz invariance of the quantum theory. Besides, applying it to the Maxwell equation  $\partial_\mu F^{\mu\nu} = 0$  one finds

$$0 = \partial_\mu \partial^\mu A^\nu - \partial_\nu (\partial_\mu A^\mu) = \partial_\mu \partial^\mu A^\nu, \quad (154)$$

which means that since  $A_\mu$  satisfies the massless Klein-Gordon equation the photon, the quantum of the electromagnetic field, has zero mass.

Once gauge invariance is fixed  $A_\mu$  is expanded in a complete basis of solutions to (154) and the canonical commutation relations are imposed

$$\widehat{A}_\mu(t, \vec{x}) = \sum_{\lambda=\pm 1} \int \frac{d^3k}{(2\pi)^3} \frac{1}{2|\vec{k}|} \left[ \epsilon_\mu(\vec{k}, \lambda) \widehat{a}(\vec{k}, \lambda) e^{-i|\vec{k}|t + i\vec{k}\cdot\vec{x}} + \epsilon_\mu(\vec{k}, \lambda)^* \widehat{a}^\dagger(\vec{k}, \lambda) e^{i|\vec{k}|t - i\vec{k}\cdot\vec{x}} \right] \quad (155)$$

where  $\lambda = \pm 1$  represent the helicity of the photon, and  $\epsilon_\mu(\vec{k}, \lambda)$  are solutions to the equations of motion with well defined momentum and helicity. Because of (153) the polarization vectors have to be orthogonal to  $k_\mu$

$$k^\mu \epsilon_\mu(\vec{k}, \lambda) = k^\mu \epsilon_\mu(\vec{k}, \lambda)^* = 0. \quad (156)$$

The canonical commutation relations imply that

$$\begin{aligned} [\widehat{a}(\vec{k}, \lambda), \widehat{a}^\dagger(\vec{k}', \lambda')] &= (2\pi)^3 (2|\vec{k}|) \delta(\vec{k} - \vec{k}') \delta_{\lambda\lambda'} \\ [\widehat{a}(\vec{k}, \lambda), \widehat{a}(\vec{k}', \lambda')] &= [\widehat{a}^\dagger(\vec{k}, \lambda), \widehat{a}^\dagger(\vec{k}', \lambda')] = 0. \end{aligned} \quad (157)$$

Therefore  $\widehat{a}(\vec{k}, \lambda)$ ,  $\widehat{a}^\dagger(\vec{k}, \lambda)$  form a set of creation-annihilation operators for photons with momentum  $\vec{k}$  and helicity  $\lambda$ .

Behind the simple construction presented above there are a number of subtleties related with gauge invariance. In particular the gauge freedom seem to introduce states in the Hilbert space with negative

probability. A careful analysis shows that when gauge invariance is properly handled these spurious states decouple from physical states and can be eliminated. The details can be found in standard textbooks [1]-[11].

**Coupling gauge fields to matter.** Once we know how to quantize the electromagnetic field we consider theories containing electrically charged particles, for example electrons. To couple the Dirac Lagrangian to electromagnetism we use as guiding principle what we learned about the Schrödinger equation for a charged particle. There we saw that the gauge ambiguity of the electromagnetic potential is compensated with a U(1) phase shift in the wave function. In the case of the Dirac equation we know that the Lagrangian is invariant under  $\psi \rightarrow e^{ie\varepsilon}\psi$ , with  $\varepsilon$  a constant. However this invariance is broken as soon as one identifies  $\varepsilon$  with the gauge transformation parameter of the electromagnetic field which depends on the position.

Looking at the Dirac Lagrangian (117) it is easy to see that in order to promote the global U(1) symmetry into a local one,  $\psi \rightarrow e^{-ie\varepsilon(x)}\psi$ , it suffices to replace the ordinary derivative  $\partial_\mu$  by a covariant one  $D_\mu$  satisfying

$$D_\mu \left[ e^{-ie\varepsilon(x)}\psi \right] = e^{-ie\varepsilon(x)}D_\mu\psi. \quad (158)$$

This covariant derivative can be constructed in terms of the gauge potential  $A_\mu$  as

$$D_\mu = \partial_\mu + ieA_\mu. \quad (159)$$

The Lagrangian of a spin- $\frac{1}{2}$  field coupled to electromagnetism is written as

$$\mathcal{L}_{\text{QED}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \bar{\psi}(i\not{D} - m)\psi, \quad (160)$$

invariant under the gauge transformations

$$\psi \longrightarrow e^{-ie\varepsilon(x)}\psi, \quad A_\mu \longrightarrow A_\mu + \partial_\mu\varepsilon(x). \quad (161)$$

Unlike the theories we have seen so far, the Lagrangian (160) describe an interacting theory. By plugging (159) into the Lagrangian we find that the interaction between fermions and photons to be

$$\mathcal{L}_{\text{QED}}^{(\text{int})} = -eA_\mu \bar{\psi}\gamma^\mu\psi. \quad (162)$$

As advertised above, in the Dirac theory the electric current four-vector is given by  $j^\mu = e\bar{\psi}\gamma^\mu\psi$ .

The quantization of interacting field theories poses new problems that we did not meet in the case of the free theories. In particular in most cases it is not possible to solve the theory exactly. When this happens the physical observables have to be computed in perturbation theory in powers of the coupling constant. An added problem appears when computing quantum corrections to the classical result, since in that case the computation of observables are plagued with infinities that should be taken care of. We will go back to this problem in section 8.

**Nonabelian gauge theories.** Quantum electrodynamics (QED) is the simplest example of a gauge theory coupled to matter based in the abelian gauge symmetry of local U(1) phase rotations. However, it is possible also to construct gauge theories based on nonabelian groups. Actually, our knowledge of the strong and weak interactions is based on the use of such nonabelian generalizations of QED.

Let us consider a gauge group  $G$  with generators  $T^a$ ,  $a = 1, \dots, \dim G$  satisfying the Lie algebra<sup>6</sup>

$$[T^a, T^b] = if^{abc}T^c. \quad (163)$$

---

<sup>6</sup>Some basics facts about Lie groups have been summarized in Appendix A.



A gauge field taking values on the Lie algebra of  $\mathcal{G}$  can be introduced  $A_\mu \equiv A_\mu^a T^a$  which transforms under a gauge transformations as

$$A_\mu \longrightarrow -\frac{1}{ig} U \partial_\mu U^{-1} + U A_\mu U^{-1}, \quad U = e^{i\chi^a(x) T^a}, \quad (164)$$

where  $g$  is the coupling constant. The associated field strength is defined as

$$F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + g f^{abc} A_\mu^b A_\nu^c. \quad (165)$$

Notice that this definition of the  $F_{\mu\nu}^a$  reduces to the one used in QED in the abelian case when  $f^{abc} = 0$ . In general, however, unlike the case of QED the field strength is not gauge invariant. In terms of  $F_{\mu\nu} = F_{\mu\nu}^a T^a$  it transforms as

$$F_{\mu\nu} \longrightarrow U F_{\mu\nu} U^{-1}. \quad (166)$$

The coupling of matter to a nonabelian gauge field is done by introducing again a covariant derivative. For a field in a representation of  $\mathcal{G}$

$$\Phi \longrightarrow U \Phi \quad (167)$$

the covariant derivative is given by

$$D_\mu \Phi = \partial_\mu \Phi - ig A_\mu^a T^a \Phi. \quad (168)$$

With the help of this we can write a generic Lagrangian for a nonabelian gauge field coupled to scalars  $\phi$  and spinors  $\psi$  as

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu}^a F^{\mu\nu a} + i \bar{\psi} \not{D} \psi + \overline{D_\mu \phi} D^\mu \phi - \bar{\psi} [M_1(\phi) + i\gamma_5 M_2(\phi)] \psi - V(\phi). \quad (169)$$

In order to keep the theory renormalizable we have to restrict  $M_1(\phi)$  and  $M_2(\phi)$  to be at most linear in  $\phi$  whereas  $V(\phi)$  have to be at most of quartic order. The Lagrangian of the standard model is of the form (169).

#### 4.4 Understanding gauge symmetry

In classical mechanics the use of the Hamiltonian formalism starts with the replacement of generalized velocities by momenta

$$p_i \equiv \frac{\partial L}{\partial \dot{q}_i} \quad \Longrightarrow \quad \dot{q}_i = \dot{q}_i(q, p). \quad (170)$$

Most of the times there is no problem in inverting the relations  $p_i = p_i(q, \dot{q})$ . However in some systems these relations might not be invertible and result in a number of constraints of the type

$$f_a(q, p) = 0, \quad a = 1, \dots, N_1. \quad (171)$$

These systems are called degenerate or constrained [23, 24].

The presence of constraints of the type (171) makes the formulation of the Hamiltonian formalism more involved. The first problem is related to the ambiguity in defining the Hamiltonian, since the addition of any linear combination of the constraints do not modify its value. Secondly, one has to make sure that the constraints are consistent with the time evolution in the system. In the language of Poisson brackets this means that further constraints have to be imposed in the form

$$\{f_a, H\} \approx 0. \quad (172)$$

Following [23] we use the symbol  $\approx$  to indicate a “weak” equality that holds when the constraints  $f_a(q, p) = 0$  are satisfied. Notice however that since the computation of the Poisson brackets involves derivatives, the constraints can be used only after the bracket is computed. In principle the conditions (172) can give rise to a new set of constraints  $g_b(q, p) = 0$ ,  $b = 1, \dots, N_2$ . Again these constraints have to be consistent with time evolution and we have to repeat the procedure. Eventually this finishes when a set of constraints is found that do not require any further constraint to be preserved by the time evolution<sup>7</sup>.

Once we find all the constraints of a degenerate system we consider the so-called first class constraints  $\phi_a(q, p) = 0$ ,  $a = 1, \dots, M$ , which are those whose Poisson bracket vanishes weakly

$$\{\phi_a, \phi_b\} = c_{abc}\phi_c \approx 0. \quad (173)$$

The constraints that do not satisfy this condition, called second class constraints, can be eliminated by modifying the Poisson bracket [23]. Then the total Hamiltonian of the theory is defined by

$$H_T = p_i q_i - L + \sum_{a=1}^M \lambda(t) \phi_a. \quad (174)$$

What has all this to do with gauge invariance? The interesting answer is that for a singular system the first class constraints  $\phi_a$  generate gauge transformations. Indeed, because  $\{\phi_a, \phi_b\} \approx 0 \approx \{\phi_a, H\}$  the transformations

$$\begin{aligned} q_i &\longrightarrow q_i + \sum_a^M \varepsilon_a(t) \{q_i, \phi_a\}, \\ p_i &\longrightarrow p_i + \sum_a^M \varepsilon_a(t) \{p_i, \phi_a\} \end{aligned} \quad (175)$$

leave invariant the state of the system. This ambiguity in the description of the system in terms of the generalized coordinates and momenta can be traced back to the equations of motion in Lagrangian language. Writing them in the form

$$\frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j} \ddot{q}_j = - \frac{\partial^2 L}{\partial \dot{q}_i \partial q_j} \dot{q}_j + \frac{\partial L}{\partial q_i}, \quad (176)$$

we find that order to determine the accelerations in terms of the positions and velocities the matrix  $\frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j}$  has to be invertible. However, the existence of constraints (171) precisely implies that the determinant of this matrix vanishes and therefore the time evolution is not uniquely determined in terms of the initial conditions.

Let us apply this to Maxwell electrodynamics described by the Lagrangian

$$L = -\frac{1}{4} \int d^3x F_{\mu\nu} F^{\mu\nu}. \quad (177)$$

The generalized momentum conjugate to  $A_\mu$  is given by

$$\pi^\mu = \frac{\delta L}{\delta(\partial_0 A_\mu)} = F^{0\mu}. \quad (178)$$

In particular for the time component we find the constraint  $\pi^0 = 0$ . The Hamiltonian is given by

$$H = \int d^3x [\pi^\mu \partial_0 A_\mu - \mathcal{L}] = \int d^3x \left[ \frac{1}{2} (\vec{E}^2 + \vec{B}^2) + \pi^0 \partial_0 A_0 + A_0 \vec{\nabla} \cdot \vec{E} \right]. \quad (179)$$

<sup>7</sup>In principle it is also possible that the procedure finishes because some kind of inconsistent identity is found. In this case the system itself is inconsistent as it is the case with the Lagrangian  $L(q, \dot{q}) = q$ .

Requiring the consistency of the constraint  $\pi^0 = 0$  we find a second constraint

$$\{\pi^0, H\} \approx \partial_0 \pi^0 + \vec{\nabla} \cdot \vec{E} = 0. \quad (180)$$

Together with the first constraint  $\pi^0 = 0$  this one implies Gauss' law  $\vec{\nabla} \cdot \vec{E} = 0$ . These two constraints have vanishing Poisson bracket and therefore they are first class. Therefore the total Hamiltonian is given by

$$H_T = H + \int d^3x \left[ \lambda_1(x) \pi^0 + \lambda_2(x) \vec{\nabla} \cdot \vec{E} \right], \quad (181)$$

where we have absorbed  $A_0$  in the definition of the arbitrary functions  $\lambda_1(x)$  and  $\lambda_2(x)$ . Actually, we can fix part of the ambiguity taking  $\lambda_1 = 0$ . Notice that, because  $A_0$  has been included in the multipliers, fixing  $\lambda_1$  amounts to fixing the value of  $A_0$  and therefore it is equivalent to taking a temporal gauge. In this case the Hamiltonian is

$$H_T = \int d^3x \left[ \frac{1}{2} (\vec{E}^2 + \vec{B}^2) + \varepsilon(x) \vec{\nabla} \cdot \vec{E} \right] \quad (182)$$

and we are left just with Gauss' law as the only constraint. Using the canonical commutation relations

$$\{A_i(t, \vec{x}), E_j(t, \vec{x}')\} = \delta_{ij} \delta(\vec{x} - \vec{x}') \quad (183)$$

we find that the remaining gauge transformations are generated by Gauss' law

$$\delta A_i = \{A_i, \int d^3x' \varepsilon \vec{\nabla} \cdot \vec{E}\} = \partial_i \varepsilon, \quad (184)$$

while leaving  $A_0$  invariant, so for consistency with the general gauge transformations the function  $\varepsilon(x)$  should be independent of time. Notice that the constraint  $\vec{\nabla} \cdot \vec{E} = 0$  can be implemented by demanding  $\vec{\nabla} \cdot \vec{A} = 0$  which reduces the three degrees of freedom of  $\vec{A}$  to the two physical degrees of freedom of the photon.

So much for the classical analysis. In the quantum theory the constraint  $\vec{\nabla} \cdot \vec{E} = 0$  has to be imposed on the physical states  $|\text{phys}\rangle$ . This is done by defining the following unitary operator on the Hilbert space

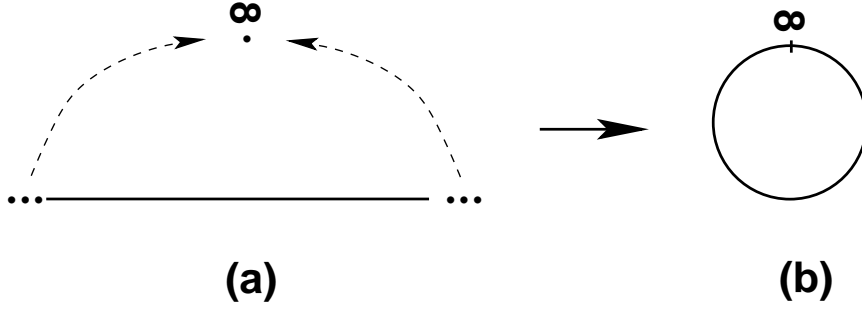
$$\mathcal{U}(\varepsilon) \equiv \exp \left( i \int d^3x \varepsilon(\vec{x}) \vec{\nabla} \cdot \vec{E} \right). \quad (185)$$

By definition, physical states should not change when a gauge transformation is performed. This is implemented by requiring that the operator  $\mathcal{U}(\varepsilon)$  acts trivially on a physical state

$$\mathcal{U}(\varepsilon) |\text{phys}\rangle = |\text{phys}\rangle \quad \Longrightarrow \quad (\vec{\nabla} \cdot \vec{E}) |\text{phys}\rangle = 0. \quad (186)$$

In the presence of charge density  $\rho$ , the condition that physical states are annihilated by Gauss' law changes to  $(\vec{\nabla} \cdot \vec{E} - \rho) |\text{phys}\rangle = 0$ .

The role of gauge transformations in the quantum theory is very illuminating in understanding the real rôle of gauge invariance [25]. As we have learned, the existence of a gauge symmetry in a theory reflects a degree of redundancy in the description of physical states in terms of the degrees of freedom appearing in the Lagrangian. In Classical Mechanics, for example, the state of a system is usually determined by the value of the canonical coordinates  $(q_i, p_i)$ . We know, however, that this is not the case for constrained Hamiltonian systems where the transformations generated by the first class constraints change the value of  $q_i$  and  $p_i$  without changing the physical state. In the case of Maxwell theory for every physical configuration determined by the gauge invariant quantities  $\vec{E}, \vec{B}$  there is an infinite number of possible values of the vector potential that are related by gauge transformations  $\delta A_\mu = \partial_\mu \varepsilon$ .



**Fig. 9:** Compactification of the real line (a) into the circumference  $S^1$  (b) by adding the point at infinity.

In the quantum theory this means that the Hilbert space of physical states is defined as the result of identifying all states related by the operator  $\mathcal{U}(\varepsilon)$  with any gauge function  $\varepsilon(x)$  into a single physical state  $|\text{phys}\rangle$ . In other words, each physical state corresponds to a whole orbit of states that are transformed among themselves by gauge transformations.

This explains the necessity of gauge fixing. In order to avoid the redundancy in the states a further condition can be given that selects one single state on each orbit. In the case of Maxwell electrodynamics the conditions  $A_0 = 0$ ,  $\vec{\nabla} \cdot \vec{A} = 0$  selects a value of the gauge potential among all possible ones giving the same value for the electric and magnetic fields.

Since states have to be identified by gauge transformations the topology of the gauge group plays an important physical rôle. To illustrate the point let us first deal with a toy model of a U(1) gauge theory in 1+1 dimensions. Later we will be more general. In the Hamiltonian formalism gauge transformations  $g(\vec{x})$  are functions defined on  $\mathbb{R}$  with values on the gauge group U(1)

$$g : \mathbb{R} \longrightarrow U(1). \quad (187)$$

We assume that  $g(x)$  is regular at infinity. In this case we can add to the real line  $\mathbb{R}$  the point at infinity to compactify it into the circumference  $S^1$  (see Fig. 9). Once this is done  $g(x)$  are functions defined on  $S^1$  with values on  $U(1) = S^1$  that can be parametrized as

$$g : S^1 \longrightarrow U(1), \quad g(x) = e^{i\alpha(x)}, \quad (188)$$

with  $x \in [0, 2\pi]$ .

Because  $S^1$  does have a nontrivial topology,  $g(x)$  can be divided into topological sectors. These sectors are labelled by an integer number  $n \in \mathbb{Z}$  and are defined by

$$\alpha(2\pi) = \alpha(0) + 2\pi n. \quad (189)$$

Geometrically  $n$  gives the number of times that the spatial  $S^1$  winds around the  $S^1$  defining the gauge group U(1). This winding number can be written in a more sophisticated way as

$$\oint_{S^1} g(x)^{-1} dg(x) = 2\pi n, \quad (190)$$

where the integral is along the spatial  $S^1$ .

In  $\mathbb{R}^3$  a similar situation happens with the gauge group<sup>8</sup> SU(2). If we demand  $g(\vec{x}) \in \text{SU}(2)$  to be regular at infinity  $|\vec{x}| \rightarrow \infty$  we can compactify  $\mathbb{R}^3$  into a three-dimensional sphere  $S^3$ , exactly as we did in 1+1 dimensions. On the other hand, the function  $g(\vec{x})$  can be written as

$$g(\vec{x}) = a^0(x)\mathbf{1} + \vec{a}(x) \cdot \vec{\sigma} \quad (191)$$

<sup>8</sup>Although we present for simplicity only the case of SU(2), similar arguments apply to any simple group.

and the conditions  $g(x)^\dagger g(x) = \mathbf{1}$ ,  $\det g = 1$  implies that  $(a^0)^2 + \vec{a}^2 = 1$ . Therefore  $SU(2)$  is a three-dimensional sphere and  $g(x)$  defines a function

$$g : S^3 \longrightarrow S^3. \quad (192)$$

As it was the case in 1+1 dimensions here the gauge transformations  $g(x)$  are also divided into topological sectors labelled this time by the winding number

$$n = \frac{1}{24\pi^2} \int_{S^3} d^3x \epsilon_{ijk} \text{Tr} [(g^{-1}\partial_i g) (g^{-1}\partial_j g) (g^{-1}\partial_k g)] \in \mathbb{Z}. \quad (193)$$

In the two cases analyzed we find that due to the nontrivial topology of the gauge group manifold the gauge transformations are divided into different sectors labelled by an integer  $n$ . Gauge transformations with different values of  $n$  cannot be smoothly deformed into each other. The sector with  $n = 0$  corresponds to those gauge transformations that can be connected with the identity.

Now we can be a bit more formal. Let us consider a gauge theory in 3+1 dimensions with gauge group  $G$  and let us denote by  $\mathcal{G}$  the set of all gauge transformations  $\mathcal{G} = \{g : S^3 \rightarrow G\}$ . At the same time we define  $\mathcal{G}_0$  as the set of transformations in  $\mathcal{G}$  that can be smoothly deformed into the identity. Our theory will have topological sectors if

$$\mathcal{G}/\mathcal{G}_0 \neq \mathbf{1}. \quad (194)$$

In the case of the electromagnetism we have seen that Gauss' law annihilates physical states. For a nonabelian theory the analysis is similar and leads to the condition

$$\mathcal{U}(g_0)|\text{phys}\rangle \equiv \exp \left[ i \int d^3x \chi^a(\vec{x}) \vec{\nabla} \cdot \vec{E}^a \right] |\text{phys}\rangle = |\text{phys}\rangle, \quad (195)$$

where  $g_0(\vec{x}) = e^{i\chi^a(\vec{x})T^a}$  is in the connected component of the identity  $\mathcal{G}_0$ . The important point to realize here is that only the elements of  $\mathcal{G}_0$  can be written as exponentials of the infinitesimal generators. Since this generators annihilate the physical states this implies that  $\mathcal{U}(g_0)|\text{phys}\rangle = |\text{phys}\rangle$  only when  $g_0 \in \mathcal{G}_0$ .

What happens then with the other topological sectors? If  $g \in \mathcal{G}/\mathcal{G}_0$  there is still a unitary operator  $\mathcal{U}(g)$  that realizes gauge transformations on the Hilbert space of the theory. However since  $g$  is not in the connected component of the identity, it cannot be written as the exponential of Gauss' law. Still gauge invariance is preserved if  $\mathcal{U}(g)$  only changes the overall global phase of the physical states. For example, if  $g_1$  is a gauge transformation with winding number  $n = 1$

$$\mathcal{U}(g_1)|\text{phys}\rangle = e^{i\theta}|\text{phys}\rangle. \quad (196)$$

It is easy to convince oneself that all transformations with winding number  $n = 1$  have the same value of  $\theta$  modulo  $2\pi$ . This can be shown by noticing that if  $g(\vec{x})$  has winding number  $n = 1$  then  $g(\vec{x})^{-1}$  has opposite winding number  $n = -1$ . Since the winding number is additive, given two transformations  $g_1, g_2$  with winding number 1,  $g_1^{-1}g_2$  has winding number  $n = 0$ . This implies that

$$|\text{phys}\rangle = \mathcal{U}(g_1^{-1}g_2)|\text{phys}\rangle = \mathcal{U}(g_1)^\dagger \mathcal{U}(g_2)|\text{phys}\rangle = e^{i(\theta_2 - \theta_1)}|\text{phys}\rangle \quad (197)$$

and we conclude that  $\theta_1 = \theta_2 \pmod{2\pi}$ . Once we know this it is straightforward to conclude that a gauge transformation  $g_n(\vec{x})$  with winding number  $n$  has the following action on physical states

$$\mathcal{U}(g_n)|\text{phys}\rangle = e^{in\theta}|\text{phys}\rangle, \quad n \in \mathbb{Z}. \quad (198)$$

To find a physical interpretation of this result we are going to look for similar things in other physical situations. One of them is borrowed from condensed matter physics and refers to the quantum

states of electrons in the periodic potential produced by the ion lattice in a solid. For simplicity we discuss the one-dimensional case where the minima of the potential are separated by a distance  $a$ . When the barrier between consecutive degenerate vacua is high enough we can neglect tunneling between different vacua and consider the ground state  $|na\rangle$  of the potential near the minimum located at  $x = na$  ( $n \in \mathbb{Z}$ ) as possible vacua of the theory. This vacuum state is, however, not invariant under lattice translations

$$e^{ia\hat{P}}|na\rangle = |(n+1)a\rangle. \quad (199)$$

However, it is possible to define a new vacuum state

$$|k\rangle = \sum_{n \in \mathbb{Z}} e^{-ikna} |na\rangle, \quad (200)$$

which under  $e^{ia\hat{P}}$  transforms by a global phase

$$e^{ia\hat{P}}|k\rangle = \sum_{n \in \mathbb{Z}} e^{-ikna} |(n+1)a\rangle = e^{ika} |k\rangle. \quad (201)$$

This ground state is labelled by the momentum  $k$  and corresponds to the Bloch wave function.

This looks very much the same as what we found for nonabelian gauge theories. The vacuum state labelled by  $\theta$  plays a rôle similar to the Bloch wave function for the periodic potential with the identification of  $\theta$  with the momentum  $k$ . To make this analogy more precise let us write the Hamiltonian for nonabelian gauge theories

$$H = \frac{1}{2} \int d^3x \left( \vec{\pi}_a \cdot \vec{\pi}_a + \vec{B}_a \cdot \vec{B}_a \right) = \frac{1}{2} \int d^3x \left( \vec{E}_a \cdot \vec{E}_a + \vec{B}_a \cdot \vec{B}_a \right), \quad (202)$$

where we have used the expression of the canonical momenta  $\pi_a^i$  and we assume that the Gauss' law constraint is satisfied. Looking at this Hamiltonian we can interpret the first term within the brackets as the kinetic energy  $T = \frac{1}{2} \vec{\pi}_a \cdot \vec{\pi}_a$  and the second term as the potential energy  $V = \frac{1}{2} \vec{B}_a \cdot \vec{B}_a$ . Since  $V \geq 0$  we can identify the vacua of the theory as those  $\vec{A}$  for which  $V = 0$ , modulo gauge transformations. This happens wherever  $\vec{A}$  is a pure gauge. However, since we know that the gauge transformations are labelled by the winding number we can have an infinite number of vacua which cannot be continuously connected with one another using trivial gauge transformations. Taking a representative gauge transformation  $g_n(\vec{x})$  in the sector with winding number  $n$ , these vacua will be associated with the gauge potentials

$$\vec{A} = -\frac{1}{ig} g_n(\vec{x}) \vec{\nabla} g_n(\vec{x})^{-1}, \quad (203)$$

modulo topologically trivial gauge transformations. Therefore the theory is characterized by an infinite number of vacua  $|n\rangle$  labelled by the winding number. These vacua are not gauge invariant. Indeed, a gauge transformation with  $n = 1$  will change the winding number of the vacua in one unit

$$\mathcal{U}(g_1)|n\rangle = |n+1\rangle. \quad (204)$$

Nevertheless a gauge invariant vacuum can be defined as

$$|\theta\rangle = \sum_{n \in \mathbb{Z}} e^{-in\theta} |n\rangle, \quad \text{with } \theta \in \mathbb{R} \quad (205)$$

satisfying

$$\mathcal{U}(g_1)|\theta\rangle = e^{i\theta} |\theta\rangle. \quad (206)$$

We have concluded that the nontrivial topology of the gauge group have very important physical consequences for the quantum theory. In particular it implies an ambiguity in the definition of the vacuum. Actually, this can also be seen in a Lagrangian analysis. In constructing the Lagrangian for the nonabelian version of Maxwell theory we only consider the term  $F_{\mu\nu}^a F^{\mu\nu a}$ . However this is not the only Lorentz and gauge invariant term that contains just two derivatives. We can write the more general Lagrangian

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^a F^{\mu\nu a} - \frac{\theta g^2}{32\pi^2} F_{\mu\nu}^a \tilde{F}^{\mu\nu a}, \quad (207)$$

where  $\tilde{F}_{\mu\nu}^a$  is the dual of the field strength defined by

$$\tilde{F}_{\mu\nu}^a = \frac{1}{2}\epsilon_{\mu\nu\sigma\lambda} F^{\sigma\lambda}. \quad (208)$$

The extra term in (207), proportional to  $\vec{E}^a \cdot \vec{B}^a$ , is actually a total derivative and does not change the equations of motion or the quantum perturbation theory. Nevertheless it has several important physical consequences. One of them is that it violates both parity  $P$  and the combination of charge conjugation and parity  $CP$ . This means that since strong interactions are described by a nonabelian gauge theory with group  $SU(3)$  there is an extra source of  $CP$  violation which puts a strong bound on the value of  $\theta$ . One of the consequences of a term like (207) in the QCD Lagrangian is a nonvanishing electric dipole moment for the neutron [26]. The fact that this is not observed impose a very strong bound on the value of the  $\theta$ -parameter

$$|\theta| < 10^{-9} \quad (209)$$

From a theoretical point of view it is still to be fully understood why  $\theta$  either vanishes or has a very small value.

Finally, the  $\theta$ -vacuum structure of gauge theories that we found in the Hamiltonian formalism can be also obtained using path integral techniques from the Lagrangian (207). The second term in Eq. (207) gives then a contribution that depends on the winding number of the corresponding gauge configuration.

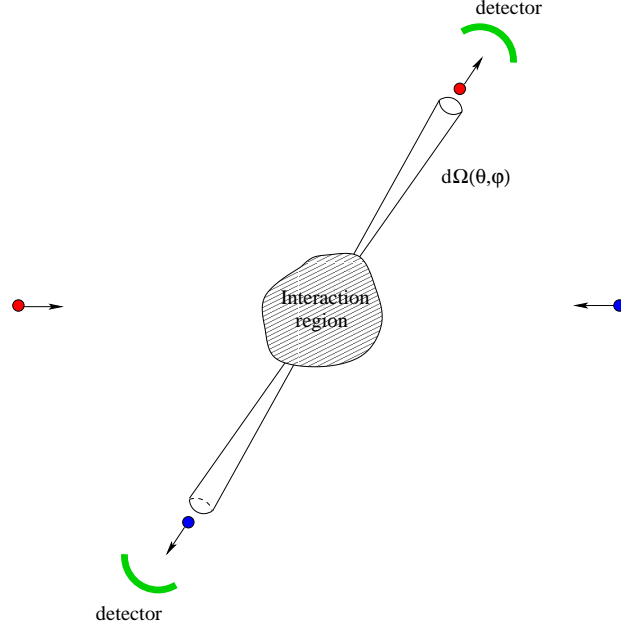
## 5 Towards computational rules: Feynman diagrams

As the basic tool to describe the physics of elementary particles, the final aim of quantum field theory is the calculation of observables. Most of the information we have about the physics of subatomic particles comes from scattering experiments. Typically, these experiments consist of arranging two or more particles to collide with a certain energy and to setup an array of detectors, sufficiently far away from the region where the collision takes place, that register the outgoing products of the collision and their momenta (together with other relevant quantum numbers).

Next we discuss how these cross sections can be computed from quantum mechanical amplitudes and how these amplitudes themselves can be evaluated in perturbative quantum field theory. We keep our discussion rather heuristic and avoid technical details that can be found in standard texts [2]- [11]. The techniques described will be illustrated with the calculation of the cross section for Compton scattering at low energies.

### 5.1 Cross sections and S-matrix amplitudes

In order to fix ideas let us consider the simplest case of a collision experiment where two particles collide to produce again two particles in the final state. The aim of such an experiments is a direct measurement of the number of particles per unit time  $\frac{dN}{dt}(\theta, \varphi)$  registered by the detector flying within a solid angle  $d\Omega$  in the direction specified by the polar angles  $\theta, \varphi$  (see Fig. 10). On general grounds we know that



**Fig. 10:** Schematic setup of a two-to-two-particles single scattering event in the center of mass reference frame.

this quantity has to be proportional to the flux of incoming particles<sup>9</sup>,  $f_{\text{in}}$ . The proportionality constant defines the differential cross section

$$\frac{dN}{dt}(\theta, \varphi) = f_{\text{in}} \frac{d\sigma}{d\Omega}(\theta, \varphi). \quad (210)$$

In natural units  $f_{\text{in}}$  has dimensions of  $(\text{length})^{-3}$ , and then the differential cross section has dimensions of  $(\text{length})^2$ . It depends, apart from the direction  $(\theta, \varphi)$ , on the parameters of the collision (energy, impact parameter, etc.) as well as on the masses and spins of the incoming particles.

Differential cross sections measure the angular distribution of the products of the collision. It is also physically interesting to quantify how effective the interaction between the particles is to produce a nontrivial dispersion. This is measured by the total cross section, which is obtained by integrating the differential cross section over all directions

$$\sigma = \int_{-1}^1 d(\cos \theta) \int_0^{2\pi} d\varphi \frac{d\sigma}{d\Omega}(\theta, \varphi). \quad (211)$$

To get some physical intuition of the meaning of the total cross section we can think of the classical scattering of a point particle off a sphere of radius  $R$ . The particle undergoes a collision only when the impact parameter is smaller than the radius of the sphere and a calculation of the total cross section yields  $\sigma = \pi R^2$ . This is precisely the cross area that the sphere presents to incoming particles.

In Quantum Mechanics in general and in quantum field theory in particular the starting point for the calculation of cross sections is the probability amplitude for the corresponding process. In a scattering experiment one prepares a system with a given number of particles with definite momenta  $\vec{p}_1, \dots, \vec{p}_n$ . In the Heisenberg picture this is described by a time independent state labelled by the incoming momenta of the particles (to keep things simple we consider spinless particles) that we denote by

$$|\vec{p}_1, \dots, \vec{p}_n; \text{in}\rangle. \quad (212)$$

<sup>9</sup>This is defined as the number of particles that enter the interaction region per unit time and per unit area perpendicular to the direction of the beam.



On the other hand, as a result of the scattering experiment a number  $k$  of particles with momenta  $\vec{p}'_1, \dots, \vec{p}'_k$  are detected. Thus, the system is now in the “out” Heisenberg picture state

$$|\vec{p}'_1, \dots, \vec{p}'_k; \text{out}\rangle \quad (213)$$

labelled by the momenta of the particles detected at late times. The probability amplitude of detecting  $k$  particles in the final state with momenta  $\vec{p}'_1, \dots, \vec{p}'_k$  in the collision of  $n$  particles with initial momenta  $\vec{p}_1, \dots, \vec{p}_n$  defines the  $S$ -matrix amplitude

$$S(\text{in} \rightarrow \text{out}) = \langle \vec{p}'_1, \dots, \vec{p}'_k; \text{out} | \vec{p}_1, \dots, \vec{p}_n; \text{in} \rangle. \quad (214)$$

It is very important to keep in mind that both the (212) and (213) are time-independent states in the Hilbert space of a very complicated interacting theory. However, since both at early and late times the incoming and outgoing particles are well apart from each other, the “in” and “out” states can be thought as two states  $|\vec{p}_1, \dots, \vec{p}_n\rangle$  and  $|\vec{p}'_1, \dots, \vec{p}'_k\rangle$  of the Fock space of the corresponding free theory in which the coupling constants are zero. Then, the overlaps (214) can be written in terms of the matrix elements of an  $S$ -matrix operator  $\hat{S}$  acting on the free Fock space

$$\langle \vec{p}'_1, \dots, \vec{p}'_k; \text{out} | \vec{p}_1, \dots, \vec{p}_n; \text{in} \rangle = \langle \vec{p}'_1, \dots, \vec{p}'_k | \hat{S} | \vec{p}_1, \dots, \vec{p}_n \rangle. \quad (215)$$

The operator  $\hat{S}$  is unitary,  $\hat{S}^\dagger = \hat{S}^{-1}$ , and its matrix elements are analytic in the external momenta.

In any scattering experiment there is the possibility that the particles do not interact at all and the system is left in the same initial state. Then it is useful to write the  $S$ -matrix operator as

$$\hat{S} = \mathbf{1} + i\hat{T}, \quad (216)$$

where  $\mathbf{1}$  represents the identity operator. In this way, all nontrivial interactions are encoded in the matrix elements of the  $T$ -operator  $\langle \vec{p}'_1, \dots, \vec{p}'_k | i\hat{T} | \vec{p}_1, \dots, \vec{p}_n \rangle$ . Since momentum has to be conserved, a global delta function can be factored out from these matrix elements to define the invariant scattering amplitude  $i\mathcal{M}$

$$\langle \vec{p}'_1, \dots, \vec{p}'_k | i\hat{T} | \vec{p}_1, \dots, \vec{p}_n \rangle = (2\pi)^4 \delta^{(4)} \left( \sum_{\text{initial}} p_i - \sum_{\text{final}} p'_f \right) i\mathcal{M}(\vec{p}_1, \dots, \vec{p}_n; \vec{p}'_1, \dots, \vec{p}'_k) \quad (217)$$

Total and differential cross sections can be now computed from the invariant amplitudes. Here we consider the most common situation in which two particles with momenta  $\vec{p}_1$  and  $\vec{p}_2$  collide to produce a number of particles in the final state with momenta  $\vec{p}'_i$ . In this case the total cross section is given by

$$\sigma = \frac{1}{(2\omega_{p_1})(2\omega_{p_2})|\vec{v}_{12}|} \int \left[ \prod_{\text{final states}} \frac{d^3 p'_i}{(2\pi)^3} \frac{1}{2\omega_{p'_i}} \right] |\mathcal{M}_{i \rightarrow f}|^2 (2\pi)^4 \delta^{(4)} \left( p_1 + p_2 - \sum_{\text{final states}} p'_i \right), \quad (218)$$

where  $\vec{v}_{12}$  is the relative velocity of the two scattering particles. The corresponding differential cross section can be computed by dropping the integration over the directions of the final momenta. We will use this expression later in Section 5.3 to evaluate the cross section of Compton scattering.

We seen how particle cross sections are determined by the invariant amplitude for the corresponding process, i.e.  $S$ -matrix amplitudes. In general, in quantum field theory it is not possible to compute exactly these amplitudes. However, in many physical situations it can be argued that interactions are weak enough to allow for a perturbative evaluation. In what follows we will describe how  $S$ -matrix elements can be computed in perturbation theory using Feynman diagrams and rules. These are very convenient bookkeeping techniques allowing both to keep track of all contributions to a process at a given order in perturbation theory, and computing the different contributions.

## 5.2 Feynman rules

The basic quantities to be computed in quantum field theory are vacuum expectation values of products of the operators of the theory. Particularly useful are time-ordered Green functions,

$$\langle \Omega | T \left[ \mathcal{O}_1(x_1) \dots \mathcal{O}_n(x_n) \right] | \Omega \rangle, \quad (219)$$

where  $|\Omega\rangle$  is the the ground state of the theory and the time ordered product is defined

$$T \left[ \mathcal{O}_i(x) \mathcal{O}_j(y) \right] = \theta(x^0 - y^0) \mathcal{O}_i(x) \mathcal{O}_j(y) + \theta(y^0 - x^0) \mathcal{O}_j(y) \mathcal{O}_i(x). \quad (220)$$

The generalization to products with more than two operators is straightforward: operators are always multiplied in time order, those evaluated at earlier times always to the right. The interest of these kind of correlation functions lies in the fact that they can be related to  $S$ -matrix amplitudes through the so-called reduction formula. To keep our discussion as simple as possible we will not derived it or even write it down in full detail. Its form for different theories can be found in any textbook. Here it suffices to say that the reduction formula simply states that any  $S$ -matrix amplitude can be written in terms of the Fourier transform of a time-ordered correlation function. Morally speaking

$$\begin{aligned} & \langle \vec{p}'_1, \dots, \vec{p}'_m; \text{out} | \vec{p}_1, \dots, \vec{p}_n; \text{in} \rangle \\ & \quad \Downarrow \\ & \int d^4x_1 \dots \int d^4y_n \langle \Omega | T \left[ \phi(x_1)^\dagger \dots \phi(x_m)^\dagger \phi(y_1) \dots \phi(y_n) \right] | \Omega \rangle e^{ip'_1 \cdot x_1} \dots e^{-ip_n \cdot y_n}, \end{aligned} \quad (221)$$

where  $\phi(x)$  is the field whose elementary excitations are the particles involved in the scattering.

The reduction formula reduces the problem of computing  $S$ -matrix amplitudes to that of evaluating time-ordered correlation functions of field operators. These quantities are easy to compute exactly in the free theory. For an interacting theory the situation is more complicated, however. Using path integrals, the vacuum expectation value of the time-ordered product of a number of operators can be expressed as

$$\langle \Omega | T \left[ \mathcal{O}_1(x_1) \dots \mathcal{O}_n(x_n) \right] | \Omega \rangle = \frac{\int \mathcal{D}\phi \mathcal{D}\phi^\dagger \mathcal{O}_1(x_1) \dots \mathcal{O}_n(x_n) e^{iS[\phi, \phi^\dagger]}}{\int \mathcal{D}\phi \mathcal{D}\phi^\dagger e^{iS[\phi, \phi^\dagger]}}. \quad (222)$$

For an theory with interactions, neither the path integral in the numerator or in the denominator is Gaussian and they cannot be calculated exactly. However, Eq. (222) is still very useful. The action  $S[\phi, \phi^\dagger]$  can be split into the free (quadratic) piece and the interaction part

$$S[\phi, \phi^\dagger] = S_0[\phi, \phi^\dagger] + S_{\text{int}}[\phi, \phi^\dagger]. \quad (223)$$

All dependence in the coupling constants of the theory comes from the second piece. Expanding now  $\exp[iS_{\text{int}}]$  in power series of the coupling constant we find that each term in the series expansion of both the numerator and the denominator has the structure

$$\int \mathcal{D}\phi \mathcal{D}\phi^\dagger \left[ \dots \right] e^{iS_0[\phi, \phi^\dagger]}, \quad (224)$$

where “...” denotes certain monomial of fields. The important point is that now the integration measure only involves the free action, and the path integral in (224) is Gaussian and therefore can be computed exactly. The same conclusion can be reached using the operator formalism. In this case the correlation function (219) can be expressed in terms of correlation functions of operators in the interaction picture. The advantage of using this picture is that the fields satisfy the free equations of motion and therefore

can be expanded in creation-annihilation operators. The correlations functions are then easily computed using Wick's theorem.

Putting together all the previous ingredients we can calculate  $S$ -matrix amplitudes in a perturbative series in the coupling constants of the field theory. This can be done using Feynman diagrams and rules, a very economical way to compute each term in the perturbative expansion of the  $S$ -matrix amplitude for a given process. We will not detail the the construction of Feynman rules but just present them heuristically.

For the sake of concreteness we focus on the case of QED first. Going back to Eq. (160) we expand the covariant derivative to write the action

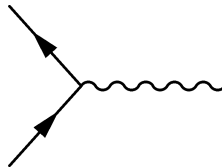
$$S_{\text{QED}} = \int d^4x \left[ -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \bar{\psi}(i\cancel{D} - m)\psi + e\bar{\psi}\gamma^\mu\psi A_\mu \right]. \quad (225)$$

The action contains two types of particles, photons and fermions, that we represent by straight and wavy lines respectively



The arrow in the fermion line does not represent the direction of the momentum but the flux of (negative) charge. This distinguishes particles from antiparticles: if the fermion propagates from left to right (i.e. in the direction of the charge flux) it represents a particle, whereas when it does from right to left it corresponds to an antiparticle. Photons are not charged and therefore wavy lines do not have orientation.

Next we turn to the interaction part of the action containing a photon field, a spinor and its conjugate. In a Feynman diagram this corresponds to the vertex



Now, in order to compute an  $S$ -matrix amplitude to a given order in the coupling constant  $e$  for a process with certain number of incoming and outgoing asymptotic states one only has to draw all possible diagrams with as many vertices as the order in perturbation theory, and the corresponding number and type of external legs. It is very important to keep in mind that in joining the fermion lines among the different building blocks of the diagram one has to respect their orientation. This reflects the conservation of the electric charge. In addition one should only consider diagrams that are topologically non-equivalent, i.e. that they cannot be smoothly deformed into one another keeping the external legs fixed<sup>10</sup>.

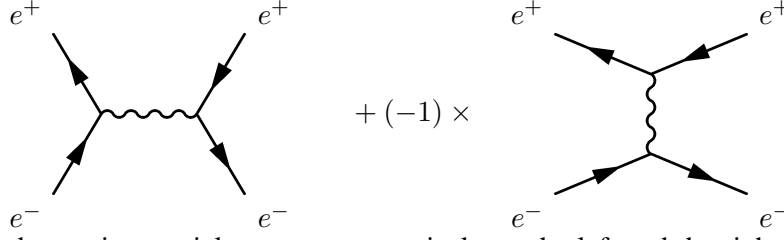
To show in a practical way how Feynman diagrams are drawn, we consider Bhabha scattering, i.e. the elastic dispersion of an electron and a positron:

$$e^+ + e^- \longrightarrow e^+ + e^-.$$

Our problem is to compute the  $S$ -matrix amplitude to the leading order in the electric charge. Because the QED vertex contains a photon line and our process does not have photons either in the initial or the

<sup>10</sup>From the point of view of the operator formalism, the requirement of considering only diagrams that are topologically nonequivalent comes from the fact that each diagram represents a certain Wick contraction in the correlation function of interaction-picture operators.

final states we find that drawing a Feynman diagram requires at least two vertices. In fact, the leading contribution is of order  $e^2$  and comes from the following two diagrams, each containing two vertices:



Incoming and outgoing particles appear respectively on the left and the right of this diagram. Notice how the identification of electrons and positrons is done comparing the direction of the charge flux with the direction of propagation. For electrons the flux of charges goes in the direction of propagation, whereas for positrons the two directions are opposite. These are the only two diagrams that can be drawn at this order in perturbation theory. It is important to include a relative minus sign between the two contributions. To understand the origin of this sign we have to remember that in the operator formalism Feynman diagrams are just a way to encode a particular Wick contraction of field operators in the interaction picture. The factor of  $-1$  reflects the relative sign in Wick contractions represented by the two diagrams, due to the fermionic character of the Dirac field.

We have learned how to draw Feynman diagrams in QED. Now one needs to compute the contribution of each one to the corresponding amplitude using the so-called Feynman rules. The idea is simple: given a diagram, each of its building blocks (vertices as well as external and internal lines) has an associated contribution that allows the calculation of the corresponding diagram. In the case of QED in the Feynman gauge, we have the following correspondence for vertices and internal propagators:

$$\begin{aligned}
 \alpha \longrightarrow \beta &\implies \left( \frac{i}{\not{p} - m + i\varepsilon} \right)_{\beta\alpha} \\
 \mu \text{ wavy } \nu &\implies \frac{-i\eta_{\mu\nu}}{p^2 + i\varepsilon} \\
 \begin{array}{c} \beta \\ \swarrow \\ \text{vertex} \\ \searrow \\ \alpha \end{array} \text{ wavy } \mu &\implies -ie\gamma_{\beta\alpha}^{\mu} (2\pi)^4 \delta^{(4)}(p_1 + p_2 + p_3).
 \end{aligned}$$

A change in the gauge would reflect in an extra piece in the photon propagator. The delta function implementing conservation of momenta is written using the convention that all momenta are entering the vertex. In addition, one has to perform an integration over all momenta running in internal lines with the measure

$$\int \frac{d^d p}{(2\pi)^4}, \tag{226}$$

and introduce a factor of  $-1$  for each fermion loop in the diagram<sup>11</sup>.

<sup>11</sup>The contribution of each diagram comes also multiplied by a degeneracy factor that takes into account in how many ways a given Wick contraction can be done. In QED, however, these factors are equal to 1 for many diagrams.

In fact, some of the integrations over internal momenta can actually be done using the delta function at the vertices, leaving just a global delta function implementing the total momentum conservation in the diagram [cf. Eq. (217)]. It is even possible that all integrations can be eliminated in this way. This is the case when we have tree level diagrams, i.e. those without closed loops. In the case of diagrams with loops there will be as many remaining integrations as the number of independent loops in the diagram.

The need to perform integrations over internal momenta in loop diagrams has important consequences in Quantum Field Theory. The reason is that in many cases the resulting integrals are ill-defined, i.e. are divergent either at small or large values of the loop momenta. In the first case one speaks of *infrared divergences* and usually they cancel once all contributions to a given process are added together. More profound, however, are the divergences appearing at large internal momenta. These *ultraviolet divergences* cannot be cancelled and have to be dealt through the renormalization procedure. We will discuss this problem in some detail in Section 8.

Were we computing time-ordered (amputated) correlation function of operators, this would be all. However, in the case of  $S$ -matrix amplitudes this is not the whole story. In addition to the previous rules here one needs to attach contributions also to the external legs in the diagram. These are the wave functions of the corresponding asymptotic states containing information about the spin and momenta of the incoming and outgoing particles. In the case of QED these contributions are:

Incoming fermion:  $\alpha \longrightarrow \text{circle with diagonal lines} \implies u_\alpha(\vec{p}, s)$

Incoming antifermion:  $\alpha \longleftarrow \text{circle with diagonal lines} \implies \bar{v}_\alpha(\vec{p}, s)$

Outgoing fermion:  $\text{circle with diagonal lines} \longrightarrow \alpha \implies \bar{u}_\alpha(\vec{p}, s)$

Outgoing antifermion:  $\text{circle with diagonal lines} \longleftarrow \alpha \implies v_\alpha(p, s)$

Incoming photon:  $\mu \text{ wavy line} \text{ circle with diagonal lines} \implies \epsilon_\mu(\vec{k}, \lambda)$

Outgoing photon:  $\text{circle with diagonal lines} \text{ wavy line} \mu \implies \epsilon_\mu(\vec{k}, \lambda)^*$

Here we have assumed that the momenta for incoming (resp. outgoing) particles are entering (resp. leaving) the diagram. It is important also to keep in mind that in the computation of  $S$ -matrix amplitudes all external states are on-shell. In Section 5.3 we illustrate the use of the Feynman rules for QED with the case of the Compton scattering.

The application of Feynman diagrams to carry out computations in perturbation theory is extremely convenient. It provides a very useful bookkeeping technique to account for all contributions to a process at a given order in the coupling constant. This does not mean that the calculation of Feynman diagrams is an easy task. The number of diagrams contributing to the process grows very fast with the order in perturbation theory and the integrals that appear in calculating loop diagrams also get very complicated. This means that, generically, the calculation of Feynman diagrams beyond the first few orders very often requires the use of computers.

Above we have illustrated the Feynman rules with the case of QED. Similar rules can be computed for other interacting quantum field theories with scalar, vector or spinor fields. In the case of the nonabelian gauge theories introduced in Section 4.3 we have:

$$\begin{aligned}
 \alpha, i \longrightarrow \beta, j &\implies \left( \frac{i}{\not{p} - m + i\varepsilon} \right)_{\beta\alpha} \delta_{ij} \\
 \mu, a \text{ (wavy) } \nu, b &\implies \frac{-i\eta_{\mu\nu}}{p^2 + i\varepsilon} \delta^{ab} \\
 \begin{array}{l} \beta, j \\ \alpha, i \end{array} \text{ (fermion lines) } \text{ meeting a wavy line } \mu, a &\implies -ig\gamma_{\beta\alpha}^{\mu} t_{ij}^a \\
 \begin{array}{l} \sigma, c \\ \nu, b \end{array} \text{ (wavy lines) } \text{ meeting a wavy line } \mu, a &\implies g f^{abc} \left[ \eta^{\mu\nu} (p_1^{\sigma} - p_2^{\sigma}) + \text{permutations} \right] \\
 \begin{array}{l} \sigma, c \\ \lambda, d \end{array} \text{ (wavy lines) } \text{ meeting a wavy line } \mu, a \\
 \begin{array}{l} \nu, b \\ \nu, b \end{array} \text{ (wavy lines) } \text{ meeting a wavy line } \mu, a &\implies -ig^2 \left[ f^{abe} f^{cde} \left( \eta^{\mu\sigma} \eta^{\nu\lambda} - \eta^{\mu\lambda} \eta^{\nu\sigma} \right) + \text{permutations} \right]
 \end{aligned}$$

It is not our aim here to give a full and detailed description of the Feynman rules for nonabelian gauge theories. It suffices to point out that, unlike the case of QED, here the gauge fields can interact

among themselves. Indeed, the three and four gauge field vertices are a consequence of the cubic and quartic terms in the action

$$S = -\frac{1}{4} \int d^4x F_{\mu\nu}^a F^{\mu\nu a}, \quad (227)$$

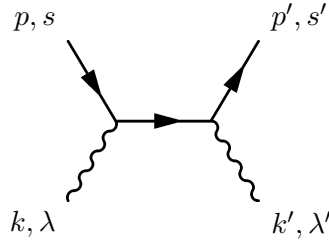
where the nonabelian gauge field strength  $F_{\mu\nu}^a$  is given in Eq. (165). The self-interaction of the non-abelian gauge fields has crucial dynamical consequences and its at the very heart of its success in describing the physics of elementary particles.

### 5.3 An example: Compton scattering

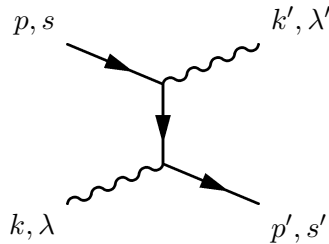
To illustrate the use of Feynman diagrams and Feynman rules we compute the cross section for the dispersion of photons by free electrons, the so-called Compton scattering:

$$\gamma(k, \lambda) + e^-(p, s) \longrightarrow \gamma(k', \lambda') + e^-(p', s').$$

In brackets we have indicated the momenta for the different particles, as well as the polarizations and spins of the incoming and outgoing photon and electrons respectively. The first step is to identify all the diagrams contributing to the process at leading order. Taking into account that the vertex of QED contains two fermion and one photon leg, it is straightforward to realize that any diagram contributing to the process at hand must contain at least two vertices. Hence the leading contribution is of order  $e^2$ . A first diagram we can draw is:



This is, however, not the only possibility. Indeed, there is a second possible diagram:



It is important to stress that these two diagrams are topologically nonequivalent, since deforming one into the other would require changing the label of the external legs. Therefore the leading  $\mathcal{O}(e^2)$  amplitude has to be computed adding the contributions from both of them.

Using the Feynman rules of QED we find

$$\begin{aligned} \text{Diagram 1} + \text{Diagram 2} &= (ie)^2 \bar{u}(\vec{p}', s') \not{\epsilon}'(\vec{k}', \lambda')^* \frac{\not{p} + \not{k} + m_e}{(p+k)^2 - m_e^2} \not{\epsilon}(\vec{k}, \lambda) u(\vec{p}, s) \\ &+ (ie)^2 \bar{u}(\vec{p}', s') \not{\epsilon}(\vec{k}, \lambda) \frac{\not{p} - \not{k}' + m_e}{(p-k')^2 - m_e^2} \not{\epsilon}'(\vec{k}', \lambda')^* u(\vec{p}, s). \end{aligned} \quad (228)$$

Because the leading order contributions only involve tree-level diagrams, there is no integration over internal momenta and therefore we are left with a purely algebraic expression for the amplitude. To get

an explicit expression we begin by simplifying the numerators. The following simple identity turns out to be very useful for this task

$$\not{a}\not{b} = -\not{b}\not{a} + 2(a \cdot b)\mathbf{1}. \quad (229)$$

Indeed, looking at the first term in Eq. (228) we have

$$\begin{aligned} (\not{p} + \not{k} + m_e)\not{\epsilon}(\vec{k}, \lambda)u(\vec{p}, s) &= -\not{\epsilon}(\vec{k}, \lambda)(\not{p} - m_e)u(\vec{p}, s) + \not{k}\not{\epsilon}(\vec{k}, \lambda)u(\vec{p}, s) \\ &+ 2p \cdot \epsilon(\vec{k}, \lambda)u(\vec{p}, s), \end{aligned} \quad (230)$$

where we have applied the identity (229) on the first term inside the parenthesis. The first term on the right-hand side of this equation vanishes identically because of Eq. (125). The expression can be further simplified if we restrict our attention to the Compton scattering at low energy when electrons are nonrelativistic. This means that all spatial momenta are much smaller than the electron mass

$$|\vec{p}|, |\vec{k}|, |\vec{p}'|, |\vec{k}'| \ll m_e. \quad (231)$$

In this approximation we have that  $p^\mu, p'^\mu \approx (m_e, \vec{0})$  and therefore

$$p \cdot \epsilon(\vec{k}, \lambda) = 0. \quad (232)$$

This follows from the absence of temporal photon polarization. Then we conclude that at low energies

$$(\not{p} + \not{k} + m_e)\not{\epsilon}(\vec{k}, \lambda)u(\vec{p}, s) = \not{k}\not{\epsilon}(\vec{k}, \lambda)u(\vec{p}, s) \quad (233)$$

and similarly for the second term in Eq. (228)

$$(\not{p} - \not{k}' + m_e)\not{\epsilon}'(\vec{k}', \lambda')^*u(\vec{p}, s) = -\not{k}'\not{\epsilon}'(\vec{k}', \lambda')^*u(\vec{p}, s). \quad (234)$$

Next, we turn to the denominators in Eq. (228). As it was explained in Section 5.2, in computing scattering amplitudes incoming and outgoing particles should have on-shell momenta,

$$p^2 = m_e^2 = p'^2 \quad \text{and} \quad k^2 = 0 = k'^2. \quad (235)$$

Then, the two denominator in Eq. (228) simplify respectively to

$$(p + k)^2 - m_e^2 = p^2 + k^2 + 2p \cdot k - m_e^2 = 2p \cdot k = 2\omega_p|\vec{k}| - 2\vec{p} \cdot \vec{k} \quad (236)$$

and

$$(p - k')^2 - m_e^2 = p^2 + k'^2 + 2p \cdot k' - m_e^2 = -2p \cdot k' = -2\omega_p|\vec{k}'| + 2\vec{p} \cdot \vec{k}'. \quad (237)$$

Working again in the low energy approximation (231) these two expressions simplify to

$$(p + k)^2 - m_e^2 \approx 2m_e|\vec{k}|, \quad (p - k')^2 - m_e^2 \approx -2m_e|\vec{k}'|. \quad (238)$$

Putting together all these expressions we find that at low energies

$$\begin{aligned} &\text{[Two Feynman diagrams: a vertex with two incoming fermion lines and two outgoing fermion lines, and a wavy photon line connecting them. The first diagram has the photon line on the left, and the second has it on the right.] +} \\ &\approx \frac{(ie)^2}{2m_e} \bar{u}(\vec{p}', s') \left[ \not{\epsilon}'(\vec{k}', \lambda')^* \frac{\not{k}}{|\vec{k}|} \epsilon(\vec{k}, \lambda) + \epsilon(\vec{k}, \lambda) \frac{\not{k}'}{|\vec{k}'|} \not{\epsilon}'(\vec{k}', \lambda')^* \right] u(\vec{p}, s). \end{aligned} \quad (239)$$



Using now again the identity (229) a number of times as well as the transversality condition of the polarization vectors (156) we end up with a handier equation

$$\begin{aligned}
 \text{Diagram 1} + \text{Diagram 2} &\approx \frac{e^2}{m_e} \left[ \epsilon(\vec{k}, \lambda) \cdot \epsilon'(\vec{k}', \lambda')^* \right] \bar{u}(\vec{p}', s') \frac{\not{k}}{|\vec{k}|} u(\vec{p}, s) \\
 &+ \frac{e^2}{2m_e} \bar{u}(\vec{p}', s') \not{\epsilon}(\vec{k}, \lambda) \not{\epsilon}'(\vec{k}', \lambda')^* \left( \frac{\not{k}}{|\vec{k}|} - \frac{\not{k}'}{|\vec{k}'|} \right) u(\vec{p}, s). \quad (240)
 \end{aligned}$$

With a little bit of effort we can show that the second term on the right-hand side vanishes. First we notice that in the low energy limit  $|\vec{k}| \approx |\vec{k}'|$ . If in addition we make use the conservation of momentum  $k - k' = p' - p$  and the identity (125)

$$\begin{aligned}
 \bar{u}(\vec{p}', s') \not{\epsilon}(\vec{k}, \lambda) \not{\epsilon}'(\vec{k}', \lambda')^* \left( \frac{\not{k}}{|\vec{k}|} - \frac{\not{k}'}{|\vec{k}'|} \right) u(\vec{p}, s) \\
 \approx \frac{1}{|\vec{k}|} \bar{u}(\vec{p}', s') \not{\epsilon}(\vec{k}, \lambda) \not{\epsilon}'(\vec{k}', \lambda')^* (\not{p}' - m_e) u(\vec{p}, s). \quad (241)
 \end{aligned}$$

Next we use the identity (229) to take the term  $(\not{p}' - m_e)$  to the right. Taking into account that in the low energy limit the electron four-momenta are orthogonal to the photon polarization vectors [see Eq. (232)] we conclude that

$$\begin{aligned}
 \bar{u}(\vec{p}', s') \not{\epsilon}(\vec{k}, \lambda) \not{\epsilon}'(\vec{k}', \lambda')^* (\not{p}' - m_e) u(\vec{p}, s) \\
 = \bar{u}(\vec{p}', s') (\not{p}' - m_e) \not{\epsilon}(\vec{k}, \lambda) \not{\epsilon}'(\vec{k}', \lambda')^* u(\vec{p}, s) = 0 \quad (242)
 \end{aligned}$$

where the last identity follows from the equation satisfied by the conjugate positive-energy spinor,  $\bar{u}(\vec{p}', s') (\not{p}' - m_e) = 0$ .

After all these lengthy manipulations we have finally arrived at the expression of the invariant amplitude for the Compton scattering at low energies

$$i\mathcal{M} = \frac{e^2}{m_e} \left[ \epsilon(\vec{k}, \lambda) \cdot \epsilon'(\vec{k}', \lambda')^* \right] \bar{u}(\vec{p}', s') \frac{\not{k}}{|\vec{k}|} u(\vec{p}, s). \quad (243)$$

The calculation of the cross section involves computing the modulus squared of this quantity. For many physical applications, however, one is interested in the dispersion of photons with a given polarization by electrons that are not polarized, i.e. whose spins are randomly distributed. In addition in many situations either we are not interested, or there is no way to measure the final polarization of the outgoing electron. This is for example the situation in cosmology, where we do not have any information about the polarization of the free electrons in the primordial plasma before or after the scattering with photons (although we have ways to measure the polarization of the scattered photons).

To describe this physical situations we have to average over initial electron polarization (since we do not know them) and sum over all possible final electron polarization (because our detector is blind to this quantum number),

$$\overline{|i\mathcal{M}|^2} = \frac{1}{2} \left( \frac{e^2}{m_e |\vec{k}|} \right)^2 \left| \epsilon(\vec{k}, \lambda) \cdot \epsilon'(\vec{k}', \lambda')^* \right|^2 \sum_{s=\pm\frac{1}{2}} \sum_{s'=\pm\frac{1}{2}} \left| \bar{u}(\vec{p}', s') \not{k} u(\vec{p}, s) \right|^2. \quad (244)$$

The factor of  $\frac{1}{2}$  comes from averaging over the two possible polarizations of the incoming electrons. The sums in this expression can be calculated without much difficulty. Expanding the absolute value explicitly

$$\sum_{s=\pm\frac{1}{2}} \sum_{s'=\pm\frac{1}{2}} \left| \bar{u}(\vec{p}', s') \not{k} u(\vec{p}, s) \right|^2 = \sum_{s=\pm\frac{1}{2}} \sum_{s'=\pm\frac{1}{2}} \left[ u(\vec{p}, s)^\dagger \not{k}^\dagger \bar{u}(\vec{p}', s')^\dagger \right] \left[ \bar{u}(\vec{p}', s') \not{k} u(\vec{p}, s) \right], \quad (245)$$

using that  $\gamma^{\mu\dagger} = \gamma^0 \gamma^\mu \gamma^0$  and after some manipulation one finds that

$$\begin{aligned} \sum_{s=\pm\frac{1}{2}} \sum_{s'=\pm\frac{1}{2}} \left| \bar{u}(\vec{p}', s') \not{k} u(\vec{p}, s) \right|^2 &= \left[ \sum_{s=\pm\frac{1}{2}} u_\alpha(\vec{p}, s) \bar{u}_\beta(\vec{p}, s) \right] (\not{k})_{\beta\sigma} \left[ \sum_{s'=\pm\frac{1}{2}} u_\sigma(\vec{p}', s') \bar{u}_\rho(\vec{p}', s') \right] (\not{k})_{\rho\alpha} \\ &= \text{Tr} \left[ (\not{p} + m_e) \not{k} (\not{p}' + m_e) \not{k} \right], \end{aligned} \quad (246)$$

where the final expression has been computed using the completeness relations in Eq. (128). The final evaluation of the trace can be done using the standard Dirac matrices identities. Here we compute it applying again the relation (229) to commute  $\not{p}'$  and  $\not{k}$ . Using that  $k^2 = 0$  and that we are working in the low energy limit we have<sup>12</sup>

$$\text{Tr} \left[ (\not{p} + m_e) \not{k} (\not{p}' + m_e) \not{k} \right] = 2(p \cdot k)(p' \cdot k) \text{Tr} \mathbf{1} \approx 8m_e^2 |\vec{k}|^2. \quad (247)$$

This gives the following value for the invariant amplitude

$$|\overline{i\mathcal{M}}|^2 = 4e^4 \left| \epsilon(\vec{k}, \lambda) \cdot \epsilon'(\vec{k}', \lambda')^* \right|^2 \quad (248)$$

Plugging  $|\overline{i\mathcal{M}}|^2$  into the formula for the differential cross section we get

$$\frac{d\sigma}{d\Omega} = \frac{1}{64\pi^2 m_e^2} |\overline{i\mathcal{M}}|^2 = \left( \frac{e^2}{4\pi m_e} \right)^2 \left| \epsilon(\vec{k}, \lambda) \cdot \epsilon'(\vec{k}', \lambda')^* \right|^2. \quad (249)$$

The prefactor of the last equation is precisely the square of the so-called classical electron radius  $r_{\text{cl}}$ . In fact, the previous differential cross section can be rewritten as

$$\frac{d\sigma}{d\Omega} = \frac{3}{8\pi} \sigma_T \left| \epsilon(\vec{k}, \lambda) \cdot \epsilon'(\vec{k}', \lambda')^* \right|^2, \quad (250)$$

where  $\sigma_T$  is the total Thomson cross section

$$\sigma_T = \frac{e^4}{6\pi m_e^2} = \frac{8\pi}{3} r_{\text{cl}}^2. \quad (251)$$

The result (250) is relevant in many areas of Physics, but its importance is paramount in the study of the cosmological microwave background (CMB). Just before recombination the universe is filled by a plasma of electrons interacting with photons via Compton scattering, with temperatures of the order of 1 keV. Electrons are then nonrelativistic ( $m_e \sim 0.5$  MeV) and the approximations leading to Eq. (250) are fully valid. Because we do not know the polarization state of the photons before being scattered by electrons we have to consider the cross section averaged over incoming photon polarizations. From Eq. (250) we see that this is proportional to

$$\frac{1}{2} \sum_{\lambda=1,2} \left| \epsilon(\vec{k}, \lambda) \cdot \epsilon'(\vec{k}', \lambda')^* \right|^2 = \left[ \frac{1}{2} \sum_{\lambda=1,2} \epsilon_i(\vec{k}, \lambda) \epsilon_j(\vec{k}, \lambda)^* \right] \epsilon_j(\vec{k}', \lambda') \epsilon_i(\vec{k}', \lambda')^*. \quad (252)$$

The sum inside the brackets can be computed using the normalization of the polarization vectors,  $|\vec{\epsilon}(\vec{k}, \lambda)|^2 = 1$ , and the transversality condition  $\vec{k} \cdot \vec{\epsilon}(\vec{k}, \lambda) = 0$

$$\frac{1}{2} \sum_{\lambda=1,2} \left| \epsilon(\vec{k}, \lambda) \cdot \epsilon'(\vec{k}', \lambda')^* \right|^2 = \frac{1}{2} \left( \delta_{ij} - \frac{k_i k_j}{|\vec{k}|^2} \right) \epsilon'_j(\vec{k}', \lambda') \epsilon'_i(\vec{k}', \lambda')^*$$

<sup>12</sup>We use also the fact that the trace of the product of an odd number of Dirac matrices is always zero.

$$= \frac{1}{2} \left[ 1 - |\vec{\ell} \cdot \vec{\epsilon}'(\vec{k}', \lambda')|^2 \right], \quad (253)$$

where  $\vec{\ell} = \frac{\vec{k}}{|\vec{k}|}$  is the unit vector in the direction of the incoming photon.

From the last equation we conclude that Thomson scattering suppresses all polarizations parallel to the direction of the incoming photon  $\vec{\ell}$ , whereas the differential cross section reaches the maximum in the plane normal to  $\vec{\ell}$ . If photons would collide with the electrons in the plasma with the same intensity from all directions, the result would be an unpolarized CMB radiation. The fact that polarization is actually measured in the CMB carries crucial information about the physics of the plasma before recombination and, as a consequence, about the very early universe (see for example [22] for a throughout discussion).

## 6 Symmetries

### 6.1 Noether's theorem

In Classical Mechanics and Classical Field Theory there is a basic result that relates symmetries and conserved charges. This is called Noether's theorem and states that for each continuous symmetry of the system there is conserved current. In its simplest version in Classical Mechanics it can be easily proved. Let us consider a Lagrangian  $L(q_i, \dot{q}_i)$  which is invariant under a transformation  $q_i(t) \rightarrow q'_i(t, \epsilon)$  labelled by a parameter  $\epsilon$ . This means that  $L(q', \dot{q}') = L(q, \dot{q})$  without using the equations of motion<sup>13</sup>. If  $\epsilon \ll 1$  we can consider an infinitesimal variation of the coordinates  $\delta_\epsilon q_i(t)$  and the invariance of the Lagrangian implies

$$0 = \delta_\epsilon L(q_i, \dot{q}_i) = \frac{\partial L}{\partial q_i} \delta_\epsilon q_i + \frac{\partial L}{\partial \dot{q}_i} \delta_\epsilon \dot{q}_i = \left[ \frac{\partial L}{\partial q_i} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} \right] \delta_\epsilon q_i + \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}_i} \delta_\epsilon q_i \right). \quad (254)$$

When  $\delta_\epsilon q_i$  is applied on a solution to the equations of motion the term inside the square brackets vanishes and we conclude that there is a conserved quantity

$$\dot{Q} = 0 \quad \text{with} \quad Q \equiv \frac{\partial L}{\partial \dot{q}_i} \delta_\epsilon q_i. \quad (255)$$

Notice that in this derivation it is crucial that the symmetry depends on a continuous parameter since otherwise the infinitesimal variation of the Lagrangian in Eq. (254) does not make sense.

In Classical Field Theory a similar result holds. Let us consider for simplicity a theory of a single field  $\phi(x)$ . We say that the variations  $\delta_\epsilon \phi$  depending on a continuous parameter  $\epsilon$  are a symmetry of the theory if, without using the equations of motion, the Lagrangian density changes by

$$\delta_\epsilon \mathcal{L} = \partial_\mu K^\mu. \quad (256)$$

If this happens then the action remains invariant and so do the equations of motion. Working out now the variation of  $\mathcal{L}$  under  $\delta_\epsilon \phi$  we find

$$\partial_\mu K^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \partial_\mu \delta_\epsilon \phi + \frac{\partial \mathcal{L}}{\partial \phi} \delta_\epsilon \phi = \partial_\mu \left( \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \delta_\epsilon \phi \right) + \left[ \frac{\partial \mathcal{L}}{\partial \phi} - \partial_\mu \left( \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \right) \right] \delta_\epsilon \phi. \quad (257)$$

If  $\phi(x)$  is a solution to the equations of motion the last terms disappears, and we find that there is a conserved current

$$\partial_\mu J^\mu = 0 \quad \text{with} \quad J^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \delta_\epsilon \phi - K^\mu. \quad (258)$$

<sup>13</sup>The following result can be also derived a more general situations where the Lagrangian changes by a total time derivative.

Actually a conserved current implies the existence of a charge

$$Q \equiv \int d^3x J^0(t, \vec{x}) \quad (259)$$

which is conserved

$$\frac{dQ}{dt} = \int d^3x \partial_0 J^0(t, \vec{x}) = - \int d^3x \partial_i J^i(t, \vec{x}) = 0, \quad (260)$$

provided the fields vanish at infinity fast enough. Moreover, the conserved charge  $Q$  is a Lorentz scalar. After canonical quantization the charge  $Q$  defined by Eq. (259) is promoted to an operator that generates the symmetry on the fields

$$\delta\phi = i[\phi, Q]. \quad (261)$$

As an example we can consider a scalar field  $\phi(x)$  which under a coordinate transformation  $x \rightarrow x'$  changes as  $\phi'(x') = \phi(x)$ . In particular performing a space-time translation  $x^{\mu'} = x^\mu + a^\mu$  we have

$$\phi'(x) - \phi(x) = -a^\mu \partial_\mu \phi + \mathcal{O}(a^2) \quad \Longrightarrow \quad \delta\phi = -a^\mu \partial_\mu \phi. \quad (262)$$

Since the Lagrangian density is also a scalar quantity, it transforms under translations as

$$\delta\mathcal{L} = -a^\mu \partial_\mu \mathcal{L}. \quad (263)$$

Therefore the corresponding conserved charge is

$$J^\mu = -\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)} a^\nu \partial_\nu \phi + a^\mu \mathcal{L} \equiv -a_\nu T^{\mu\nu}, \quad (264)$$

where we introduced the energy-momentum tensor

$$T^{\mu\nu} = \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)} \partial^\nu \phi - \eta^{\mu\nu} \mathcal{L}. \quad (265)$$

We find that associated with the invariance of the theory with respect to space-time translations there are four conserved currents defined by  $T^{\mu\nu}$  with  $\nu = 0, \dots, 3$ , each one associated with the translation along a space-time direction. These four currents form a rank-two tensor under Lorentz transformations satisfying

$$\partial_\mu T^{\mu\nu} = 0. \quad (266)$$

The associated conserved charges are given by

$$P^\nu = \int d^3x T^{0\nu} \quad (267)$$

and correspond to the total energy-momentum content of the field configuration. Therefore the energy density of the field is given by  $T^{00}$  while  $T^{0i}$  is the momentum density. In the quantum theory the  $P^\mu$  are the generators of space-time translations.

Another example of a symmetry related with a physically relevant conserved charge is the global phase invariance of the Dirac Lagrangian (117),  $\psi \rightarrow e^{i\theta}\psi$ . For small  $\theta$  this corresponds to variations  $\delta_\theta\psi = i\theta\psi$ ,  $\delta_\theta\bar{\psi} = -i\theta\bar{\psi}$  which by Noether's theorem result in the conserved charge

$$j^\mu = \bar{\psi}\gamma^\mu\psi, \quad \partial_\mu j^\mu = 0. \quad (268)$$

Thus implying the existence of a conserved charge

$$Q = \int d^3x \bar{\psi} \gamma^0 \psi = \int d^3x \psi^\dagger \psi. \quad (269)$$

In physics there are several instances of global U(1) symmetries that act as phase shifts on spinors. This is the case, for example, of the baryon and lepton number conservation in the standard model. A more familiar case is the U(1) local symmetry associated with electromagnetism. Notice that although in this case we are dealing with a local symmetry,  $\theta \rightarrow e\alpha(x)$ , the invariance of the Lagrangian holds in particular for global transformations and therefore there is a conserved current  $j^\mu = e\bar{\psi}\gamma^\mu\psi$ . In Eq. (162) we saw that the spinor is coupled to the photon field precisely through this current. Its time component is the electric charge density  $\rho$ , while the spatial components are the current density vector  $\vec{j}$ .

This analysis can be carried over also to nonabelian unitary global symmetries acting as

$$\psi_i \longrightarrow U_{ij} \psi_j, \quad U^\dagger U = \mathbf{1} \quad (270)$$

and leaving invariant the Dirac Lagrangian when we have several fermions. If we write the matrix  $U$  in terms of the hermitian group generators  $T^a$  as

$$U = \exp(i\alpha_a T^a), \quad (T^a)^\dagger = T^a, \quad (271)$$

we find the conserved current

$$j^{\mu a} = \bar{\psi}_i T_{ij}^a \gamma^\mu \psi_j, \quad \partial_\mu j^\mu = 0. \quad (272)$$

This is the case, for example of the approximate flavor symmetries in hadron physics. The simplest example is the isospin symmetry that mixes the quarks  $u$  and  $d$

$$\begin{pmatrix} u \\ d \end{pmatrix} \longrightarrow M \begin{pmatrix} u \\ d \end{pmatrix}, \quad M \in \text{SU}(2). \quad (273)$$

Since the proton is a bound state of two quarks  $u$  and one quark  $d$  while the neutron is made out of one quark  $u$  and two quarks  $d$ , this isospin symmetry reduces at low energies to the well known isospin transformations of nuclear physics that mixes protons and neutrons.

## 6.2 Symmetries in the quantum theory

We have seen that in canonical quantization the conserved charges  $Q^a$  associated to symmetries by Noether's theorem are operators implementing the symmetry at the quantum level. Since the charges are conserved they must commute with the Hamiltonian

$$[Q^a, H] = 0. \quad (274)$$

There are several possibilities in the quantum mechanical realization of a symmetry:

**Wigner-Weyl realization.** In this case the ground state of the theory  $|0\rangle$  is invariant under the symmetry. Since the symmetry is generated by  $Q^a$  this means that

$$\mathcal{U}(\alpha)|0\rangle \equiv e^{i\alpha_a Q^a}|0\rangle = |0\rangle \implies Q^a|0\rangle = 0. \quad (275)$$

At the same time the fields of the theory have to transform according to some irreducible representation of the group generated by the  $Q^a$ . From Eq. (261) it is easy to prove that

$$\mathcal{U}(\alpha)\phi_i\mathcal{U}(\alpha)^{-1} = U_{ij}(\alpha)\phi_j, \quad (276)$$

where  $U_{ij}(\alpha)$  is an element of the representation in which the field  $\phi_i$  transforms. If we consider now the quantum state associated with the operator  $\phi_i$

$$|i\rangle = \phi_i|0\rangle \quad (277)$$

we find that because of the invariance of the vacuum (275) the states  $|i\rangle$  transform in the same representation as  $\phi_i$

$$\mathcal{U}(\alpha)|i\rangle = \mathcal{U}(\alpha)\phi_i\mathcal{U}(\alpha)^{-1}\mathcal{U}(\alpha)|0\rangle = U_{ij}(\alpha)\phi_j|0\rangle = U_{ij}(\alpha)|j\rangle. \quad (278)$$

Therefore the spectrum of the theory is classified in multiplets of the symmetry group. In addition, since  $[H, \mathcal{U}(\alpha)] = 0$  all states in the same multiplet have the same energy. If we consider one-particle states, then going to the rest frame we conclude that all states in the same multiplet have exactly the same mass.

**Nambu-Goldstone realization.** In our previous discussion the result that the spectrum of the theory is classified according to multiplets of the symmetry group depended crucially on the invariance of the ground state. However this condition is not mandatory and one can relax it to consider theories where the vacuum state is not left invariant by the symmetry

$$e^{i\alpha_a Q^a}|0\rangle \neq |0\rangle \quad \implies \quad Q^a|0\rangle \neq 0. \quad (279)$$

In this case it is also said that the symmetry is spontaneously broken by the vacuum.

To illustrate the consequences of (279) we consider the example of a number scalar fields  $\varphi^i$  ( $i = 1, \dots, N$ ) whose dynamics is governed by the Lagrangian

$$\mathcal{L} = \frac{1}{2}\partial_\mu\varphi^i\partial^\mu\varphi^i - V(\varphi), \quad (280)$$

where we assume that  $V(\phi)$  is bounded from below. This theory is globally invariant under the transformations

$$\delta\varphi^i = \epsilon^a (T^a)^i_j \varphi^j, \quad (281)$$

with  $T^a$ ,  $a = 1, \dots, \frac{1}{2}N(N-1)$  the generators of the group  $\text{SO}(N)$ .

To analyze the structure of vacua of the theory we construct the Hamiltonian

$$H = \int d^3x \left[ \frac{1}{2}\pi^i\pi^i + \frac{1}{2}\vec{\nabla}\varphi^i \cdot \vec{\nabla}\varphi^i + V(\varphi) \right] \quad (282)$$

and look for the minimum of

$$\mathcal{V}(\varphi) = \int d^3x \left[ \frac{1}{2}\vec{\nabla}\varphi^i \cdot \vec{\nabla}\varphi^i + V(\varphi) \right]. \quad (283)$$

Since we are interested in finding constant field configurations,  $\vec{\nabla}\varphi = \vec{0}$  to preserve translational invariance, the vacua of the potential  $\mathcal{V}(\varphi)$  coincides with the vacua of  $V(\varphi)$ . Therefore the minima of the potential correspond to the vacuum expectation values<sup>14</sup>

$$\langle\varphi^i\rangle : \quad V(\langle\varphi^i\rangle) = 0, \quad \left. \frac{\partial V}{\partial\varphi^i} \right|_{\varphi^i=\langle\varphi^i\rangle} = 0. \quad (284)$$

We divide the generators  $T^a$  of  $\text{SO}(N)$  into two groups: Those denoted by  $H^\alpha$  ( $\alpha = 1, \dots, h$ ) that satisfy

$$(H^\alpha)^i_j \langle\varphi^j\rangle = 0. \quad (285)$$

<sup>14</sup>For simplicity we consider that the minima of  $V(\phi)$  occur at zero potential.

This means that the vacuum configuration  $\langle \varphi^i \rangle$  is left invariant by the transformation generated by  $H^\alpha$ . For this reason we call them *unbroken generators*. Notice that the commutator of two unbroken generators also annihilates the vacuum expectation value,  $[H^\alpha, H^\beta]_{ij} \langle \varphi^j \rangle = 0$ . Therefore the generators  $\{H^\alpha\}$  form a subalgebra of the algebra of the generators of  $SO(N)$ . The subgroup of the symmetry group generated by them is realized à la Wigner-Weyl.

The remaining generators  $K^A$ , with  $A = 1, \dots, \frac{1}{2}N(N-1) - h$ , by definition do not preserve the vacuum expectation value of the field

$$(K^A)_j^i \langle \varphi^j \rangle \neq 0. \quad (286)$$

These will be called the *broken generators*. Next we prove a very important result concerning the broken generators known as the Goldstone theorem: for each generator broken by the vacuum expectation value there is a massless excitation.

The mass matrix of the excitations around the vacuum  $\langle \varphi^i \rangle$  is determined by the quadratic part of the potential. Since we assumed that  $V(\langle \varphi \rangle) = 0$  and we are expanding around a minimum, the first term in the expansion of the potential  $V(\varphi)$  around the vacuum expectation values is given by

$$V(\varphi) = \left. \frac{\partial^2 V}{\partial \varphi^i \partial \varphi^j} \right|_{\varphi = \langle \varphi \rangle} (\varphi^i - \langle \varphi^i \rangle)(\varphi^j - \langle \varphi^j \rangle) + \mathcal{O}[(\varphi - \langle \varphi \rangle)^3] \quad (287)$$

and the mass matrix is:

$$M_{ij}^2 \equiv \left. \frac{\partial^2 V}{\partial \varphi^i \partial \varphi^j} \right|_{\varphi = \langle \varphi \rangle}. \quad (288)$$

In order to avoid a cumbersome notation we do not show explicitly the dependence of the mass matrix on the vacuum expectation values  $\langle \varphi^i \rangle$ .

To extract some information about the possible zero modes of the mass matrix, we write down the conditions that follow from the invariance of the potential under  $\delta \varphi^i = \epsilon^a (T^a)_j^i \varphi^j$ . At first order in  $\epsilon^a$

$$\delta V(\varphi) = \epsilon^a \frac{\partial V}{\partial \varphi^i} (T^a)_j^i \varphi^j = 0. \quad (289)$$

Differentiating this expression with respect to  $\varphi^k$  we arrive at

$$\frac{\partial^2 V}{\partial \varphi^i \partial \varphi^k} (T^a)_j^i \varphi^j + \frac{\partial V}{\partial \varphi^i} (T^a)_k^i = 0. \quad (290)$$

Now we evaluate this expression in the vacuum  $\varphi^i = \langle \varphi^i \rangle$ . Then the derivative in the second term cancels while the second derivative in the first one gives the mass matrix. Hence we find

$$M_{ik}^2 (T^a)_j^i \langle \varphi^j \rangle = 0. \quad (291)$$

Now we can write this expression for both broken and unbroken generators. For the unbroken ones, since  $(H^\alpha)_j^i \langle \varphi^j \rangle = 0$ , we find a trivial identity  $0 = 0$ . On the other hand for the broken generators we have

$$M_{ik}^2 (K^A)_j^i \langle \varphi^j \rangle = 0. \quad (292)$$

Since  $(K^A)_j^i \langle \varphi^j \rangle \neq 0$  this equation implies that the mass matrix has as many zero modes as broken generators. Therefore we have proven Goldstone's theorem: associated with each broken symmetry there is a massless mode in the theory. Here we have presented a classical proof of the theorem. In the quantum theory the proof follows the same lines as the one presented here but one has to consider the effective action containing the effects of the quantum corrections to the classical Lagrangian.

As an example to illustrate this theorem, we consider a SO(3) invariant scalar field theory with a “mexican hat” potential

$$V(\vec{\varphi}) = \frac{\lambda}{4} (\vec{\varphi}^2 - a^2)^2. \quad (293)$$

The vacua of the theory correspond to the configurations satisfying  $\langle \vec{\varphi} \rangle^2 = a^2$ . In field space this equation describes a two-dimensional sphere and each solution is just a point in that sphere. Geometrically it is easy to visualize that a given vacuum field configuration, i.e. a point in the sphere, is preserved by SO(2) rotations around the axis of the sphere that passes through that point. Hence the vacuum expectation value of the scalar field breaks the symmetry according to

$$\langle \vec{\varphi} \rangle : \quad \text{SO}(3) \longrightarrow \text{SO}(2). \quad (294)$$

Since SO(3) has three generators and SO(2) only one we see that two generators are broken and therefore there are two massless Goldstone bosons. Physically this massless modes can be thought of as corresponding to excitations along the surface of the sphere  $\langle \vec{\varphi} \rangle^2 = a^2$ .

Once a minimum of the potential has been chosen we can proceed to quantize the excitations around it. Since the vacuum only leaves invariant a SO(2) subgroup of the original SO(3) symmetry group it seems that the fact that we are expanding around a particular vacuum expectation value of the scalar field has resulted in a lost of symmetry. This is however not the case. The full quantum theory is symmetric under the whole symmetry group SO(3). This is reflected in the fact that the physical properties of the theory do not depend on the particular point of the sphere  $\langle \vec{\varphi} \rangle^2 = a^2$  that we have chosen. Different vacua are related by the full SO(3) symmetry and therefore should give the same physics.

It is very important to realize that given a theory with a vacuum determined by  $\langle \vec{\varphi} \rangle$  all other possible vacua of the theory are inaccessible in the infinite volume limit. This means that two vacuum states  $|0_1\rangle, |0_2\rangle$  corresponding to different vacuum expectation values of the scalar field are orthogonal  $\langle 0_1|0_2\rangle = 0$  and cannot be connected by any local observable  $\Phi(x)$ ,  $\langle 0_1|\Phi(x)|0_2\rangle = 0$ . Heuristically this can be understood by noticing that in the infinite volume limit switching from one vacuum into another one requires changing the vacuum expectation value of the field everywhere in space at the same time, something that cannot be done by any local operator. Notice that this is radically different to our expectations based on the Quantum Mechanics of a system with a finite number of degrees of freedom.

In High Energy Physics the typical example of a Goldstone boson is the pion, associated with the spontaneous breaking of the global chiral isospin  $\text{SU}(2)_L \times \text{SU}(2)_R$  symmetry. This symmetry acts independently in the left- and right-handed spinors as

$$\begin{pmatrix} u_{L,R} \\ d_{L,R} \end{pmatrix} \longrightarrow M_{L,R} \begin{pmatrix} u_{L,R} \\ d_{L,R} \end{pmatrix}, \quad M_{L,R} \in \text{SU}(2)_{L,R} \quad (295)$$

Presumably since the quarks are confined at low energies this symmetry is spontaneously broken down to the diagonal SU(2) acting in the same way on the left- and right-handed components of the spinors. Associated with this symmetry breaking there is a Goldstone mode which is identified as the pion. Notice, nevertheless, that the  $\text{SU}(2)_L \times \text{SU}(2)_R$  would be an exact global symmetry of the QCD Lagrangian only in the limit when the masses of the quarks are zero  $m_u, m_d \rightarrow 0$ . Since these quarks have nonzero masses the chiral symmetry is only approximate and as a consequence the corresponding Goldstone boson is not massless. That is why pions have masses, although they are the lightest particle among the hadrons.

Symmetry breaking appears also in many places in condensed matter. For example, when a solid crystallizes from a liquid the translational invariance that is present in the liquid phase is broken to a discrete group of translations that represent the crystal lattice. This symmetry breaking has Goldstone



bosons associated which are identified with phonons which are the quantum excitation modes of the vibrational degrees of freedom of the lattice.

**The Higgs mechanism.** Gauge symmetry seems to prevent a vector field from having a mass. This is obvious once we realize that a term in the Lagrangian like  $m^2 A_\mu A^\mu$  is incompatible with gauge invariance.

However certain physical situations seem to require massive vector fields. This happened for example during the 1960s in the study of weak interactions. The Glashow model gave a common description of both electromagnetic and weak interactions based on a gauge theory with group  $SU(2) \times U(1)$  but, in order to reproduce Fermi's four-fermion theory of the  $\beta$ -decay it was necessary that two of the vector fields involved would be massive. Also in condensed matter physics massive vector fields are required to describe certain systems, most notably in superconductivity.

The way out to this situation is found in the concept of spontaneous symmetry breaking discussed previously. The consistency of the quantum theory requires gauge invariance, but this invariance can be realized à la Nambu-Goldstone. When this is the case the full gauge symmetry is not explicitly present in the effective action constructed around the particular vacuum chosen by the theory. This makes possible the existence of mass terms for gauge fields without jeopardizing the consistency of the full theory, which is still invariant under the whole gauge group.

To illustrate the Higgs mechanism we study the simplest example, the Abelian Higgs model: a  $U(1)$  gauge field coupled to a self-interacting charged complex scalar field  $\Phi$  with Lagrangian

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \overline{D}_\mu \Phi D^\mu \Phi - \frac{\lambda}{4} (\overline{\Phi} \Phi - \mu^2)^2, \quad (296)$$

where the covariant derivative is given by Eq. (159). This theory is invariant under the gauge transformations

$$\Phi \rightarrow e^{i\alpha(x)} \Phi, \quad A_\mu \rightarrow A_\mu + \partial_\mu \alpha(x). \quad (297)$$

The minimum of the potential is defined by the equation  $|\Phi| = \mu$ . We have a continuum of different vacua labelled by the phase of the scalar field. None of these vacua, however, is invariant under the gauge symmetry

$$\langle \Phi \rangle = \mu e^{i\vartheta_0} \rightarrow \mu e^{i\vartheta_0 + i\alpha(x)} \quad (298)$$

and therefore the symmetry is spontaneously broken. Let us study now the theory around one of these vacua, for example  $\langle \Phi \rangle = \mu$ , by writing the field  $\Phi$  in terms of the excitations around this particular vacuum

$$\Phi(x) = \left[ \mu + \frac{1}{\sqrt{2}} \sigma(x) \right] e^{i\vartheta(x)}. \quad (299)$$

Independently of whether we are expanding around a particular vacuum for the scalar field we should keep in mind that the whole Lagrangian is still gauge invariant under (297). This means that performing a gauge transformation with parameter  $\alpha(x) = -\vartheta(x)$  we can get rid of the phase in Eq. (299). Substituting then  $\Phi(x) = \mu + \frac{1}{\sqrt{2}} \sigma(x)$  in the Lagrangian we find

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + e^2 \mu^2 A_\mu A^\mu + \frac{1}{2} \partial_\mu \sigma \partial^\mu \sigma - \frac{1}{2} \lambda \mu^2 \sigma^2 \\ &\quad - \lambda \mu \sigma^3 - \frac{\lambda}{4} \sigma^4 + e^2 \mu A_\mu A^\mu \sigma + e^2 A_\mu A^\mu \sigma^2. \end{aligned} \quad (300)$$

What are the excitation of the theory around the vacuum  $\langle \Phi \rangle = \mu$ ? First we find a massive real scalar field  $\sigma(x)$ . The important point however is that the vector field  $A_\mu$  now has a mass given by

$$m_\gamma^2 = 2e^2 \mu^2. \quad (301)$$

The remarkable thing about this way of giving a mass to the photon is that at no point we have given up gauge invariance. The symmetry is only hidden. Therefore in quantizing the theory we can still enjoy all the advantages of having a gauge theory but at the same time we have managed to generate a mass for the gauge field.

It is surprising, however, that in the Lagrangian (300) we did not find any massless mode. Since the vacuum chosen by the scalar field breaks the  $U(1)$  generator of  $U(1)$  we would have expected one massless particle from Goldstone's theorem. To understand the fate of the missing Goldstone boson we have to revisit the calculation leading to Eq. (300). Were we dealing with a global  $U(1)$  theory, the Goldstone boson would correspond to excitation of the scalar field along the valley of the potential and the phase  $\vartheta(x)$  would be the massless Goldstone boson. However we have to keep in mind that in computing the Lagrangian we managed to get rid of  $\vartheta(x)$  by shifting it into  $A_\mu$  using a gauge transformation. Actually by identifying the gauge parameter with the Goldstone excitation we have completely fixed the gauge and the Lagrangian (300) does not have any gauge symmetry left.

A massive vector field has three polarizations: two transverse ones  $\vec{k} \cdot \vec{\epsilon}(\vec{k}, \pm 1) = 0$  plus a longitudinal one  $\vec{\epsilon}_L(\vec{k}) \sim \vec{k}$ . In gauging away the massless Goldstone boson  $\vartheta(x)$  we have transformed it into the longitudinal polarization of the massive vector field. In the literature this is usually expressed saying that the Goldstone mode is “eaten up” by the longitudinal component of the gauge field. It is important to realize that in spite of the fact that the Lagrangian (300) looks pretty different from the one we started with we have not lost any degrees of freedom. We started with the two polarizations of the photon plus the two degrees of freedom associated with the real and imaginary components of the complex scalar field. After symmetry breaking we end up with the three polarizations of the massive vector field and the degree of freedom of the real scalar field  $\sigma(x)$ .

We can also understand the Higgs mechanism in the light of our discussion of gauge symmetry in section 4.4. In the Higgs mechanism the invariance of the theory under infinitesimal gauge transformations is not explicitly broken, and this implies that Gauss' law is satisfied quantum mechanically,  $\vec{\nabla} \cdot \vec{E}_a|_{\text{phys}} = 0$ . The theory remains invariant under gauge transformations in the connected component of the identity  $\mathcal{G}_0$ , the ones generated by Gauss' law. This does not pose any restriction on the possible breaking of the invariance of the theory with respect to transformations that cannot be continuously deformed to the identity. Hence in the Higgs mechanism the invariance under gauge transformation that are not in the connected component of the identity,  $\mathcal{G}/\mathcal{G}_0$ , can be broken. Let us try to put it in more precise terms. As we learned in section 4.4, in the Hamiltonian formulation of the theory finite energy gauge field configurations tend to a pure gauge at spatial infinity

$$\vec{A}_\mu(\vec{x}) \longrightarrow -\frac{1}{ig} g(\vec{x}) \vec{\nabla} g(\vec{x})^{-1}, \quad |\vec{x}| \rightarrow \infty \quad (302)$$

The set transformations  $g_0(\vec{x}) \in \mathcal{G}_0$  that tend to the identity at infinity are the ones generated by Gauss' law. However, one can also consider in general gauge transformations  $g(\vec{x})$  which, as  $|\vec{x}| \rightarrow \infty$ , approach any other element  $g \in G$ . The quotient  $\mathcal{G}_\infty \equiv \mathcal{G}/\mathcal{G}_0$  gives a copy of the gauge group at infinity. There is no reason, however, why this group should not be broken, and in general it is if the gauge symmetry is spontaneously broken. Notice that this is not a threat to the consistency of the theory. Properties like the decoupling of unphysical states are guaranteed by the fact that Gauss' law is satisfied quantum mechanically and are not affected by the breaking of  $\mathcal{G}_\infty$ .

In condensed matter physics the symmetry breaking described by the nonrelativistic version of the Abelian Higgs model can be used to characterize the onset of a superconducting phase in the BCS theory, where the complex scalar field  $\Phi$  is associated with the Cooper pairs. In this case the parameter  $\mu^2$  depends on the temperature. Above the critical temperature  $T_c$ ,  $\mu^2(T) > 0$  and there is only a symmetric vacuum  $\langle \Phi \rangle = 0$ . When, on the other hand,  $T < T_c$  then  $\mu^2(T) < 0$  and symmetry breaking takes place. The onset of a nonzero mass of the photon (301) below the critical temperature explains the Meissner effect: the magnetic fields cannot penetrate inside superconductors beyond a distance of the order  $\frac{1}{m_\gamma}$ .

The Abelian Higgs model discussed here can be regarded as a toy model of the Brout-Englert-Higgs mechanism responsible for giving mass to the  $W^\pm$  and  $Z^0$  gauge bosons in the standard model. Giving mass to these three bosons requires the introduction of a two-component complex scalar field transforming as a doublet under  $SU(2)$ . Three of its four degrees of freedom are incorporated as the longitudinal components of the three massive gauge fields, whereas the fourth one remains as a scalar propagating degree of freedom. Its elementary excitations are spin zero neutral particles known as Higgs bosons.

The Higgs boson couples to the massive gauge fields, as well as to quarks and leptons. Moreover, its coupling to the fermions is proportional to the fermion masses and therefore very weak for light fermions. This, together with the fact that Higgs production processes have large standard model backgrounds, complicates its experimental detection. After decades of searches in various experiments, a Higgs boson candidate was finally detected at the ATLAS and CMS collaborations at the Large Hadron Collider (LHC) in 2012 with a mass of approximately 125 GeV. At the time of writing, all evidences point to the fact that this new particle is indeed the so much coveted standard model Higgs.

## 7 Anomalies

So far we did not worry too much about how classical symmetries of a theory are carried over to the quantum theory. We have implicitly assumed that classical symmetries are preserved in the process of quantization, so they are also realized in the quantum theory.

This, however, does not have to be necessarily the case. Quantizing an interacting field theory is a very involved process that requires regularization and renormalization and sometimes, it does not matter how hard we try, there is no way for a classical symmetry to survive quantization. When this happens one says that the theory has an *anomaly* (for reviews see [28]). It is important to avoid here the misconception that anomalies appear due to a bad choice of the way a theory is regularized in the process of quantization. When we talk about anomalies we mean a classical symmetry that *cannot* be realized in the quantum theory, no matter how smart we are in choosing the regularization procedure.

In the following we analyze some examples of anomalies associated with global and local symmetries of the classical theory. In Section 8 we will encounter yet another example of an anomaly, this time associated with the breaking of classical scale invariance in the quantum theory.

### 7.1 Axial anomaly

Probably the best known examples of anomalies appear when we consider axial symmetries. If we consider a theory of two Weyl spinors  $u_\pm$

$$\mathcal{L} = i\bar{\psi}\not{\partial}\psi = iu_+^\dagger\sigma_+^\mu\partial_\mu u_+ + iu_-^\dagger\sigma_-^\mu\partial_\mu u_- \quad \text{with} \quad \psi = \begin{pmatrix} u_+ \\ u_- \end{pmatrix} \quad (303)$$

the Lagrangian is invariant under two types of global  $U(1)$  transformations. In the first one both helicities transform with the same phase, this is a *vector* transformation:

$$U(1)_V : u_\pm \longrightarrow e^{i\alpha}u_\pm, \quad (304)$$

whereas in the second one, the axial  $U(1)$ , the signs of the phases are different for the two chiralities

$$U(1)_A : u_\pm \longrightarrow e^{\pm i\alpha}u_\pm. \quad (305)$$

Using Noether's theorem, there are two conserved currents, a vector current

$$J_V^\mu = \bar{\psi}\gamma^\mu\psi = u_+^\dagger\sigma_+^\mu u_+ + u_-^\dagger\sigma_-^\mu u_- \quad \implies \quad \partial_\mu J_V^\mu = 0 \quad (306)$$

and an axial vector current

$$J_A^\mu = \bar{\psi} \gamma^\mu \gamma_5 \psi = u_+^\dagger \sigma_+^\mu u_+ - u_-^\dagger \sigma_-^\mu u_- \quad \Longrightarrow \quad \partial_\mu J_A^\mu = 0. \quad (307)$$

The theory described by the Lagrangian (303) can be coupled to the electromagnetic field. The resulting classical theory is still invariant under the vector and axial U(1) symmetries (304) and (305). Surprisingly, upon quantization it turns out that the conservation of the axial current (307) is spoiled by quantum effects

$$\partial_\mu J_A^\mu \sim \hbar \vec{E} \cdot \vec{B}. \quad (308)$$

To understand more clearly how this result comes about we study first a simple model in two dimensions that captures the relevant physics involved in the four-dimensional case [29]. We work in Minkowski space in two dimensions with coordinates  $(x^0, x^1) \equiv (t, x)$  and where the spatial direction is compactified to a circle  $S^1$ . In this setup we consider a fermion coupled to the electromagnetic field. Notice that since we are living in two dimensions the field strength  $F_{\mu\nu}$  only has one independent component that corresponds to the electric field along the spatial direction,  $F^{01} \equiv \mathcal{E}$  (in two dimensions there are no magnetic fields!).

To write the Lagrangian for the spinor field we need to find a representation of the algebra of  $\gamma$ -matrices

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu} \quad \text{with} \quad \eta = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (309)$$

In two dimensions the dimension of the representation of the  $\gamma$ -matrices is  $2^{\lfloor \frac{2}{2} \rfloor} = 2$ . Here take

$$\gamma^0 \equiv \sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \gamma^1 \equiv i\sigma^2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \quad (310)$$

This is a chiral representation since the matrix  $\gamma_5$  is diagonal<sup>15</sup>

$$\gamma_5 \equiv -\gamma^0 \gamma^1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (311)$$

Writing the two-component spinor  $\psi$  as

$$\psi = \begin{pmatrix} u_+ \\ u_- \end{pmatrix} \quad (312)$$

and defining as usual the projectors  $P_\pm = \frac{1}{2}(\mathbf{1} \pm \gamma_5)$  we find that the components  $u_\pm$  of  $\psi$  are respectively a right- and left-handed Weyl spinor in two dimensions.

Once we have a representation of the  $\gamma$ -matrices we can write the Dirac equation. Expressing it in terms of the components  $u_\pm$  of the Dirac spinor we find

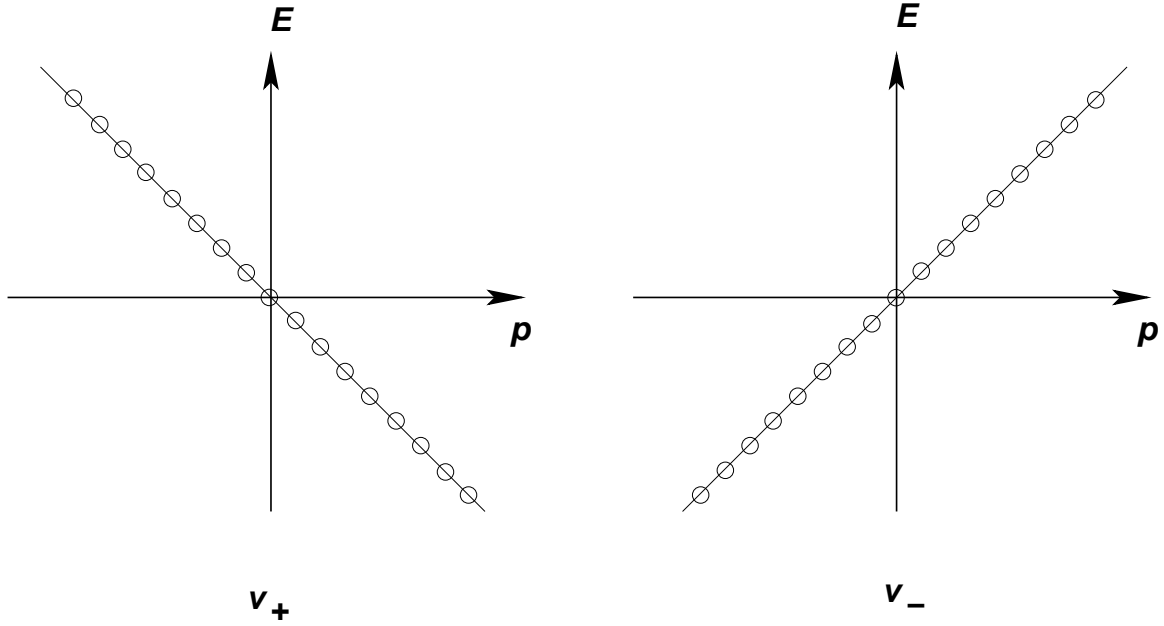
$$(\partial_0 - \partial_1)u_+ = 0, \quad (\partial_0 + \partial_1)u_- = 0. \quad (313)$$

The general solution to these equations can be immediately written as

$$u_+ = u_+(x^0 + x^1), \quad u_- = u_-(x^0 - x^1). \quad (314)$$

Hence  $u_\pm$  are two wave packets moving along the spatial dimension respectively to the left ( $u_+$ ) and to the right ( $u_-$ ). Notice that according to our convention the left-moving  $u_+$  is a right-handed spinor (positive helicity) whereas the right-moving  $u_-$  is a left-handed spinor (negative helicity).

<sup>15</sup>In any even number of dimensions  $\gamma_5$  is defined to satisfy the conditions  $\gamma_5^2 = \mathbf{1}$  and  $\{\gamma_5, \gamma^\mu\} = 0$ .



**Fig. 11:** Spectrum of the massless two-dimensional Dirac field.

If we want to interpret (313) as the wave equation for two-dimensional Weyl spinors we have the following wave functions for free particles with well defined momentum  $p^\mu = (E, p)$ .

$$u_{\pm}^{(E)}(x^0 \pm x^1) = \frac{1}{\sqrt{L}} e^{-iE(x^0 \pm x^1)} \quad \text{with} \quad p = \mp E. \quad (315)$$

As it is always the case with the Dirac equation we have both positive and negative energy solutions. For  $u_+$ , since  $E = -p$ , we see that the solutions with positive energy are those with negative momentum  $p < 0$ , whereas the negative energy solutions are plane waves with  $p > 0$ . For the left-handed spinor  $u_-$  the situation is reversed. Besides, since the spatial direction is compact with length  $L$  the momentum  $p$  is quantized according to

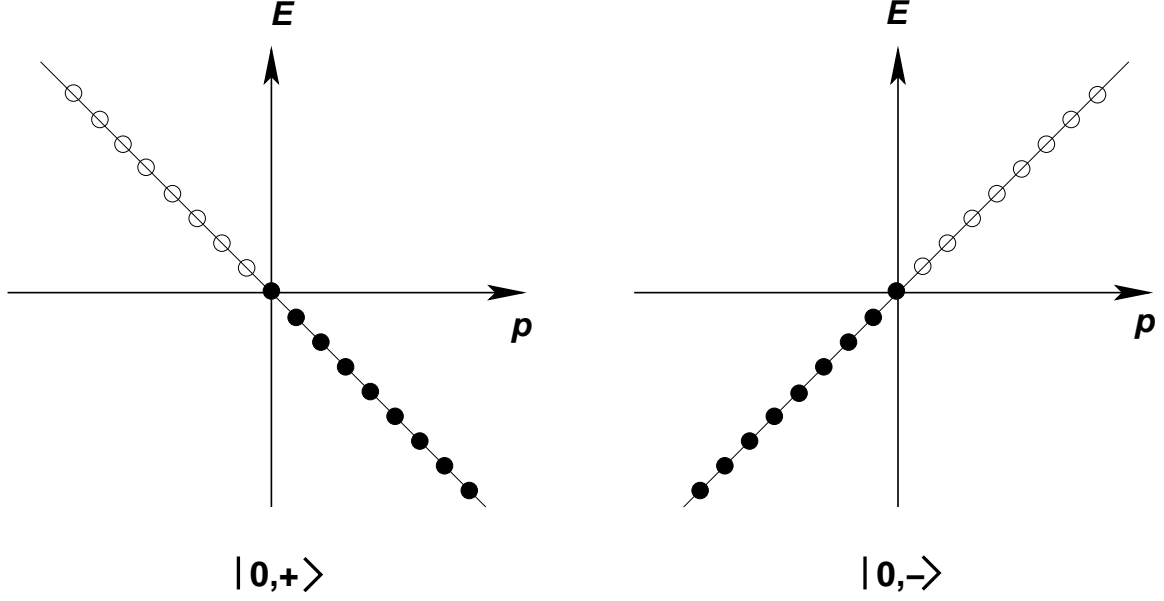
$$p = \frac{2\pi n}{L}, \quad n \in \mathbb{Z}. \quad (316)$$

The spectrum of the theory is represented in Fig. 11.

Once we have the spectrum of the theory the next step is to obtain the vacuum. As with the Dirac equation in four dimensions we fill all the states with  $E \leq 0$  (Fig. 12). Exciting of a particle in the Dirac sea produces a positive energy fermion plus a hole that is interpreted as an antiparticle. This gives us the clue on how to quantize the theory. In the expansion of the operator  $u_{\pm}$  in terms of the modes (315) we associate positive energy states with annihilation operators whereas the states with negative energy are associated with creation operators for the corresponding antiparticle

$$u_{\pm}(x) = \sum_{E>0} \left[ a_{\pm}(E) v_{\pm}^{(E)}(x) + b_{\pm}^{\dagger}(E) v_{\pm}^{(E)}(x)^* \right]. \quad (317)$$

The operator  $a_{\pm}(E)$  acting on the vacuum  $|0, \pm\rangle$  annihilates a particle with positive energy  $E$  and momentum  $\mp E$ . In the same way  $b_{\pm}^{\dagger}(E)$  creates out of the vacuum an antiparticle with positive energy  $E$  and spatial momentum  $\mp E$ . In the Dirac sea picture the operator  $b_{\pm}(E)^{\dagger}$  is originally an annihilation operator for a state of the sea with negative energy  $-E$ . As in the four-dimensional case the problem of the negative energy states is solved by interpreting annihilation operators for negative energy states as



**Fig. 12:** Vacuum of the theory.

creation operators for the corresponding antiparticle with positive energy (and vice versa). The operators appearing in the expansion of  $u_{\pm}$  in Eq. (317) satisfy the usual algebra

$$\{a_{\lambda}(E), a_{\lambda'}^{\dagger}(E')\} = \{b_{\lambda}(E), b_{\lambda'}^{\dagger}(E')\} = \delta_{E,E'}\delta_{\lambda\lambda'}, \quad (318)$$

where we have introduced the label  $\lambda, \lambda' = \pm$ . Also,  $a_{\lambda}(E), a_{\lambda'}^{\dagger}(E)$  anticommute with  $b_{\lambda'}(E'), b_{\lambda'}^{\dagger}(E')$ .

The Lagrangian of the theory

$$\mathcal{L} = iu_{+}^{\dagger}(\partial_0 + \partial_1)u_{+} + iu_{-}^{\dagger}(\partial_0 - \partial_1)u_{-} \quad (319)$$

is invariant under both  $U(1)_V$ , Eq. (304), and  $U(1)_A$ , Eq. (305). The associated Noether currents are in this case

$$J_V^{\mu} = \begin{pmatrix} u_{+}^{\dagger}u_{+} + u_{-}^{\dagger}u_{-} \\ -u_{+}^{\dagger}u_{+} + u_{-}^{\dagger}u_{-} \end{pmatrix}, \quad J_A^{\mu} = \begin{pmatrix} u_{+}^{\dagger}u_{+} - u_{-}^{\dagger}u_{-} \\ -u_{+}^{\dagger}u_{+} - u_{-}^{\dagger}u_{-} \end{pmatrix}. \quad (320)$$

The associated conserved charges are given, for the vector current by

$$Q_V = \int_0^L dx^1 \left( u_{+}^{\dagger}u_{+} + u_{-}^{\dagger}u_{-} \right) \quad (321)$$

and for the axial current

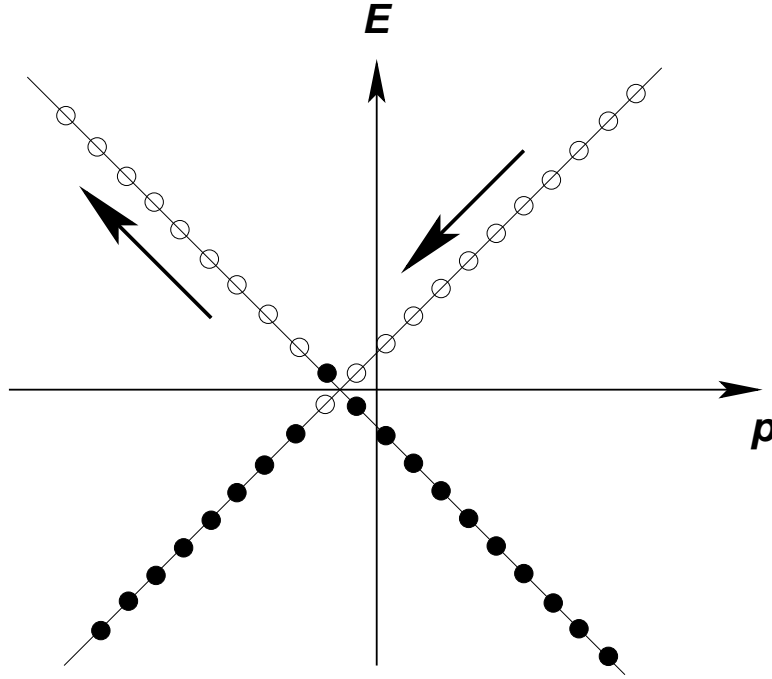
$$Q_A = \int_0^L dx^1 \left( u_{+}^{\dagger}u_{+} - u_{-}^{\dagger}u_{-} \right). \quad (322)$$

Using the orthonormality relations for the modes  $v_{\pm}^{(E)}(x)$

$$\int_0^L dx^1 v_{\pm}^{(E)}(x) v_{\pm}^{(E')}(x) = \delta_{E,E'} \quad (323)$$

we find for the conserved charges:

$$Q_V = \sum_{E>0} \left[ a_{+}^{\dagger}(E)a_{+}(E) - b_{+}^{\dagger}(E)b_{+}(E) + a_{-}^{\dagger}(E)a_{-}(E) - b_{-}^{\dagger}(E)b_{-}(E) \right],$$



**Fig. 13:** Effect of the electric field.

$$Q_A = \sum_{E>0} \left[ a_+^\dagger(E) a_+(E) - b_+^\dagger(E) b_+(E) - a_-^\dagger(E) a_-(E) + b_-^\dagger(E) b_-(E) \right]. \quad (324)$$

We see that  $Q_V$  counts the net number (particles minus antiparticles) of positive helicity states plus the net number of states with negative helicity. The axial charge, on the other hand, counts the net number of positive helicity states minus the number of negative helicity ones. In the case of the vector current we have subtracted a formally divergent vacuum contribution to the charge (the “charge of the Dirac sea”).

In the free theory there is of course no problem with the conservation of either  $Q_V$  or  $Q_A$ , since the occupation numbers do not change. What we want to study is the effect of coupling the theory to electric field  $\mathcal{E}$ . We work in the gauge  $A_0 = 0$ . Instead of solving the problem exactly we are going to simulate the electric field by adiabatically varying in a long time  $\tau_0$  the vector potential  $A_1$  from zero value to  $-\mathcal{E}\tau_0$ . From our discussion in section 4.3 we know that the effect of the electromagnetic coupling in the theory is a shift in the momentum according to

$$p \longrightarrow p - eA_1, \quad (325)$$

where  $e$  is the charge of the fermions. Since we assumed that the vector potential varies adiabatically, we can assume it to be approximately constant at each time.

Then, we have to understand what is the effect of (325) on the vacuum depicted in Fig. (12). What we find is that the two branches move as shown in Fig. (13) resulting in some of the negative energy states of the  $v_+$  branch acquiring positive energy while the same number of the empty positive energy states of the other branch  $v_-$  will become empty negative energy states. Physically this means that the external electric field  $\mathcal{E}$  creates a number of particle-antiparticle pairs out of the vacuum. Denoting by  $N \sim e\mathcal{E}$  the number of such pairs created by the electric field per unit time, the final values of the charges  $Q_V$  and  $Q_A$  are

$$\begin{aligned} Q_A(\tau_0) &= (N - 0) + (0 - N) = 0, \\ Q_V(\tau_0) &= (N - 0) - (0 - N) = 2N. \end{aligned} \quad (326)$$





from the gauge symmetry, this Lagrangian is also invariant under a global  $U(N_f)_L \times U(N_f)_R$  acting on the flavor indices and defined by

$$U(N_f)_L : \begin{cases} Q_L^f \rightarrow \sum_{f'} (U_L)_{ff'} Q_L^{f'} \\ Q_R^f \rightarrow Q_R^f \end{cases} \quad U(N_f)_R : \begin{cases} Q_L^f \rightarrow Q_L^f \\ Q_R^f \rightarrow \sum_{f'} (U_R)_{ff'} Q_R^{f'} \end{cases} \quad (332)$$

with  $U_L, U_R \in U(N_f)$ . Actually, since  $U(N) = U(1) \times SU(N)$  this global symmetry group can be written as  $SU(N_f)_L \times SU(N_f)_R \times U(1)_L \times U(1)_R$ . The abelian subgroup  $U(1)_L \times U(1)_R$  can be now decomposed into their vector  $U(1)_B$  and axial  $U(1)_A$  subgroups defined by the transformations

$$U(1)_B : \begin{cases} Q_L^f \rightarrow e^{i\alpha} Q_L^f \\ Q_R^f \rightarrow e^{i\alpha} Q_R^f \end{cases} \quad U(1)_A : \begin{cases} Q_L^f \rightarrow e^{i\alpha} Q_L^f \\ Q_R^f \rightarrow e^{-i\alpha} Q_R^f \end{cases} \quad (333)$$

According to Noether's theorem, associated with these two abelian symmetries we have two conserved currents:

$$J_V^\mu = \sum_{f=1}^{N_f} \bar{Q}^f \gamma^\mu Q^f, \quad J_A^\mu = \sum_{f=1}^{N_f} \bar{Q}^f \gamma^\mu \gamma_5 Q^f. \quad (334)$$

The conserved charge associated with vector charge  $J_V^\mu$  is actually the baryon number defined as the number of quarks minus number of antiquarks.

The nonabelian part of the global symmetry group  $SU(N_f)_L \times SU(N_f)_R$  can also be decomposed into its vector and axial subgroups,  $SU(N_f)_V \times SU(N_f)_A$ , defined by the following transformations of the quarks fields

$$SU(N_f)_V : \begin{cases} Q_L^f \rightarrow \sum_{f'} (U_L)_{ff'} Q_L^{f'} \\ Q_R^f \rightarrow \sum_{f'} (U_L)_{ff'} Q_R^{f'} \end{cases} \quad SU(N_f)_A : \begin{cases} Q_L^f \rightarrow \sum_{f'} (U_L)_{ff'} Q_L^{f'} \\ Q_R^f \rightarrow \sum_{f'} (U_R^{-1})_{ff'} Q_R^{f'} \end{cases} \quad (335)$$

Again, the application of Noether's theorem shows the existence of the following nonabelian conserved charges

$$J_V^{I\mu} \equiv \sum_{f,f'=1}^{N_f} \bar{Q}^f \gamma^\mu (T^I)_{ff'} Q^{f'}, \quad J_A^{I\mu} \equiv \sum_{f,f'=1}^{N_f} \bar{Q}^f \gamma^\mu \gamma_5 (T^I)_{ff'} Q^{f'}. \quad (336)$$

To summarize, we have shown that the initial chiral symmetry of the QCD Lagrangian (331) can be decomposed into its chiral and vector subgroups according to

$$U(N_f)_L \times U(N_f)_R = SU(N_f)_V \times SU(N_f)_A \times U(1)_B \times U(1)_A. \quad (337)$$

The question to address now is which part of the classical global symmetry is preserved by the quantum theory.

As argued in section 7.1, the conservation of the axial currents  $J_A^\mu$  and  $J_A^{a\mu}$  can in principle be spoiled due to the presence of an anomaly. In the case of the abelian axial current  $J_A^\mu$  the relevant quantity is the correlation function

$$C^{\mu\nu\sigma} \equiv \langle 0 | T \left[ J_A^\mu(x) j_{\text{gauge}}^{a\nu}(x') j_{\text{gauge}}^{b\sigma}(0) \right] | 0 \rangle = \sum_{f=1}^{N_f} \left[ \begin{array}{c} \text{Diagram} \end{array} \right]_{\text{symmetric}} \quad (338)$$

Here  $j_{\text{gauge}}^{a\mu}$  is the nonabelian conserved current coupling to the gluon field

$$j_{\text{gauge}}^{a\mu} \equiv \sum_{f=1}^{N_f} \bar{Q}^f \gamma^\mu \tau^a Q^f, \quad (339)$$

where, to avoid confusion with the generators of the global symmetry we have denoted by  $\tau^a$  the generators of the gauge group  $SU(N_c)$ . The anomaly can be read now from  $\partial_\mu C^{\mu\nu\sigma}$ . If we impose Bose symmetry with respect to the interchange of the two outgoing gluons and gauge invariance of the whole expression,  $\partial_\nu C^{\mu\nu\sigma} = 0 = \partial_\sigma C^{\mu\nu\sigma}$ , we find that the axial abelian global current has an anomaly given by<sup>16</sup>

$$\partial_\mu J_A^\mu = -\frac{g^2 N_f}{32\pi^2} \varepsilon^{\mu\nu\sigma\lambda} F_{\mu\nu}^a F_{\sigma\lambda}^a. \quad (340)$$

In the case of the nonabelian axial global symmetry  $SU(N_f)_A$  the calculation of the anomaly is made as above. The result, however, is quite different since in this case we conclude that the nonabelian axial current  $J_A^{a\mu}$  is not anomalous. This can be easily seen by noticing that associated with the axial current vertex we have a generator  $T^I$  of  $SU(N_f)$ , whereas for the two gluon vertices we have the generators  $\tau^a$  of the gauge group  $SU(N_c)$ . Therefore, the triangle diagram is proportional to the group-theoretic factor

$$\left[ \begin{array}{c} \text{Diagram: Triangle with } J_A^{I\mu} \text{ on the left, } Q^f \text{ on the top and bottom, and } g \text{ on the right.} \\ \text{symmetric} \end{array} \right] \sim \text{tr } T^I \text{tr } \{\tau^a, \tau^b\} = 0 \quad (341)$$

which vanishes because the generators of  $SU(N_f)$  are traceless.

From here we would conclude that the nonabelian axial symmetry  $SU(N_f)_A$  is nonanomalous. However this is not the whole story since quarks are charged particles that also couple to photons. Hence there is a second potential source of an anomaly coming from the one-loop triangle diagram coupling  $J_A^{I\mu}$  to two photons

$$\langle 0|T [J_A^{I\mu}(x) j_{\text{em}}^\nu(x') j_{\text{em}}^\sigma(0)] |0\rangle = \sum_{f=1}^{N_f} \left[ \begin{array}{c} \text{Diagram: Triangle with } J_A^{I\mu} \text{ on the left, } Q^f \text{ on the top and bottom, and } \gamma \text{ on the right.} \\ \text{symmetric} \end{array} \right] \quad (342)$$

where  $j_{\text{em}}^\mu$  is the electromagnetic current

$$j_{\text{em}}^\mu = \sum_{f=1}^{N_f} q_f \bar{Q}^f \gamma^\mu Q^f, \quad (343)$$

with  $q_f$  the electric charge of the  $f$ -th quark flavor. A calculation of the diagram in (342) shows the existence of an Adler-Bell-Jackiw anomaly given by

$$\partial_\mu J_A^{I\mu} = -\frac{N_c}{16\pi^2} \left[ \sum_{f=1}^{N_f} (T^I)_{ff} q_f^2 \right] \varepsilon^{\mu\nu\sigma\lambda} F_{\mu\nu} F_{\sigma\lambda}, \quad (344)$$

<sup>16</sup>The normalization of the generators  $T^I$  of the global  $SU(N_f)$  is given by  $\text{tr}(T^I T^J) = \frac{1}{2} \delta^{IJ}$ .

where  $F_{\mu\nu}$  is the field strength of the electromagnetic field coupling to the quarks. The only chance for the anomaly to cancel is that the factor between brackets in this equation be identically zero.

Before proceeding let us summarize the results found so far. Because of the presence of anomalies the axial part of the global chiral symmetry,  $SU(N_f)_A$  and  $U(1)_A$  are not realized quantum mechanically in general. We found that  $U(1)_A$  is always affected by an anomaly. However, because the right-hand side of the anomaly equation (340) is a total derivative, the anomalous character of  $J_A^\mu$  does not explain the absence of  $U(1)_A$  multiplets in the hadron spectrum, since a new current can be constructed which is conserved. In addition, the nonexistence of candidates for a Goldstone boson associated with the right quantum numbers indicates that  $U(1)_A$  is not spontaneously broken either, so it has to be explicitly broken somehow. This is the so-called  $U(1)$ -problem which was solved by 't Hooft [33], who showed how the contribution of quantum transitions between vacua with topologically nontrivial gauge field configurations (instantons) results in an explicit breaking of this symmetry.

Due to the dynamics of the  $SU(N_c)$  gauge theory the axial nonabelian symmetry is spontaneously broken due to the presence at low energies of a vacuum expectation value for the fermion bilinear  $\bar{Q}^f Q^f$

$$\langle 0 | \bar{Q}^f Q^f | 0 \rangle \neq 0 \quad (\text{No summation in } f!). \quad (345)$$

This nonvanishing vacuum expectation value for the quark bilinear actually breaks chiral invariance spontaneously to the vector subgroup  $SU(N_f)_V$ , so the only subgroup of the original global symmetry that is realized by the full theory at low energy is

$$SU(N_f)_L \times SU(N_f)_R \longrightarrow SU(N_f)_V \times U(1)_B. \quad (346)$$

Associated with this breaking a Goldstone boson should appear with the quantum numbers of the broken nonabelian current. For example, in the case of QCD the Goldstone bosons associated with the spontaneously symmetry breaking induced by the vacuum expectation values  $\langle \bar{u}u \rangle$ ,  $\langle \bar{d}d \rangle$  and  $\langle (\bar{u}d - \bar{d}u) \rangle$  have been identified as the pions  $\pi^0$ ,  $\pi^\pm$ . These bosons are not exactly massless because of the nonvanishing mass of the  $u$  and  $d$  quarks. Since the global chiral symmetry is already slightly broken by mass terms in the Lagrangian, the associated Goldstone bosons also have masses although they are very light compared to the masses of other hadrons.

In order to have a better physical understanding of the role of anomalies in the physics of strong interactions we particularize now our analysis of the case of real QCD. Since the  $u$  and  $d$  quarks are much lighter than the other four flavors, QCD at low energies can be well described by including only these two flavors and ignoring heavier quarks. In this approximation, from our previous discussion we know that the low energy global symmetry of the theory is  $SU(2)_V \times U(1)_B$ , where now the vector group  $SU(2)_V$  is the well-known isospin symmetry. The axial  $U(1)_A$  current is anomalous due to Eq. (340) with  $N_f = 2$ . In the case of the nonabelian axial symmetry  $SU(2)_A$ , taking into account that  $q_u = \frac{2}{3}e$  and  $q_d = -\frac{1}{3}e$  and that the three generators of  $SU(2)$  can be written in terms of the Pauli matrices as  $T^K = \frac{1}{2}\sigma^K$  we find

$$\sum_{f=u,d} (T^1)_{ff} q_f^2 = \sum_{f=u,d} (T^2)_{ff} q_f^2 = 0, \quad \sum_{f=u,d} (T^3)_{ff} q_f^2 = \frac{e^2}{6}. \quad (347)$$

Therefore  $J_A^{3\mu}$  is anomalous.

Physically, the anomaly in the axial current  $J_A^{3\mu}$  has an important consequence. In the quark model, the wave function of the neutral pion  $\pi^0$  is given in terms of those for the  $u$  and  $d$  quark by

$$|\pi^0\rangle = \frac{1}{\sqrt{2}} (|\bar{u}\rangle|u\rangle - |\bar{d}\rangle|d\rangle). \quad (348)$$

The isospin quantum numbers of  $|\pi^0\rangle$  are those of the generator  $T^3$ . Actually the analogy goes further since  $\partial_\mu J_A^{3\mu}$  is the operator creating a pion  $\pi^0$  out of the vacuum

$$|\pi^0\rangle \sim \partial_\mu J_A^{3\mu}|0\rangle. \quad (349)$$

This leads to the physical interpretation of the triangle diagram (342) with  $J_A^{3\mu}$  as the one loop contribution to the decay of a neutral pion into two photons

$$\pi^0 \longrightarrow 2\gamma. \quad (350)$$

This is an interesting piece of physics. In 1967 Sutherland and Veltman [34] presented a calculation, using current algebra techniques, according to which the decay of the pion into two photons should be suppressed. This however contradicted the experimental evidence that showed the existence of such a decay. The way out to this paradox, as pointed out in [30], is the axial anomaly. What happens is that the current algebra analysis overlooks the ambiguities associated with the regularization of divergences in quantum field theory. A QED evaluation of the triangle diagram leads to a divergent integral that has to be regularized somehow. It is in this process that the Adler-Bell-Jackiw axial anomaly appears resulting in a nonvanishing value for the  $\pi^0 \rightarrow 2\gamma$  amplitude<sup>17</sup>.

The existence of anomalies associated with global currents does not necessarily mean difficulties for the theory. On the contrary, as we saw in the case of the axial anomaly it is its existence what allows for a solution of the Sutherland-Veltman paradox and an explanation of the electromagnetic decay of the pion. The situation, however, is very different if we deal with local symmetries. A quantum mechanical violation of gauge symmetry leads to all kinds of problems, from lack of renormalizability to nondecoupling of negative norm states. This is because the presence of an anomaly in the theory implies that the Gauss' law constraint  $\vec{\nabla} \cdot \vec{E}_a = \rho_a$  cannot be consistently implemented in the quantum theory. As a consequence states that classically are eliminated by the gauge symmetry become propagating fields in the quantum theory, thus spoiling the consistency of the theory.

Anomalies in a gauge symmetry can be expected only in chiral theories where left and right-handed fermions transform in different representations of the gauge group. Physically, the most interesting example of such theories is the electroweak sector of the standard model where, for example, left handed fermions transform as doublets under SU(2) whereas right-handed fermions are singlets. On the other hand, QCD is free of gauge anomalies since both left- and right-handed quarks transform in the fundamental representation of SU(3).

We consider the Lagrangian

$$\mathcal{L} = -\frac{1}{4}F^{a\mu\nu}F_{\mu\nu}^a + i \sum_{i=1}^{N_+} \bar{\psi}_+^i \mathcal{D}^{(+)} \psi_+^i + i \sum_{j=1}^{N_-} \bar{\psi}_-^j \mathcal{D}^{(-)} \psi_-^j, \quad (351)$$

where the chiral fermions  $\psi_\pm^i$  transform according to the representations  $\tau_{i,\pm}^a$  of the gauge group  $G$  ( $a = 1, \dots, \dim G$ ). The covariant derivatives  $D_\mu^{(\pm)}$  are then defined by

$$D_\mu^{(\pm)} \psi_\pm^i = \partial_\mu \psi_\pm^i + ig A_\mu^K \tau_{i,\pm}^K \psi_\pm^i. \quad (352)$$

As for global symmetries, anomalies in the gauge symmetry appear in the triangle diagram with one axial and two vector gauge current vertices

$$\langle 0|T [j_A^{a\mu}(x)j_V^{b\nu}(x')j_V^{c\sigma}(0)]|0\rangle = \left[ \text{triangle diagram} \right]_{\text{symmetric}} \quad (353)$$

<sup>17</sup>An early computation of the triangle diagram for the electromagnetic decay of the pion was made by Steinberger in [31].

where gauge vector and axial currents  $j_V^{a\mu}$ ,  $j_A^{a\mu}$  are given by

$$\begin{aligned} j_V^{a\mu} &= \sum_{i=1}^{N_+} \bar{\psi}_+^i \tau_+^a \gamma^\mu \psi_+^i + \sum_{j=1}^{N_-} \bar{\psi}_-^j \tau_-^a \gamma^\mu \psi_-^j, \\ j_A^{a\mu} &= \sum_{i=1}^{N_+} \bar{\psi}_+^i \tau_+^a \gamma^\mu \psi_+^i - \sum_{j=1}^{N_-} \bar{\psi}_-^j \tau_-^a \gamma^\mu \psi_-^j. \end{aligned} \quad (354)$$

Luckily, we do not have to compute the whole diagram in order to find an anomaly cancellation condition, it is enough if we calculate the overall group theoretical factor. In the case of the diagram in Eq. (353) for every fermion species running in the loop this factor is equal to

$$\text{tr} \left[ \tau_{i,\pm}^a \{ \tau_{i,\pm}^b, \tau_{i,\pm}^c \} \right], \quad (355)$$

where the sign  $\pm$  corresponds respectively to the generators of the representation of the gauge group for the left and right-handed fermions. Hence the anomaly cancellation condition reads

$$\sum_{i=1}^{N_+} \text{tr} \left[ \tau_{i,+}^a \{ \tau_{i,+}^b, \tau_{i,+}^c \} \right] - \sum_{j=1}^{N_-} \text{tr} \left[ \tau_{j,-}^a \{ \tau_{j,-}^b, \tau_{j,-}^c \} \right] = 0. \quad (356)$$

Knowing this we can proceed to check the anomaly cancellation in the standard model  $SU(3) \times SU(2) \times U(1)$ . Left handed fermions (both leptons and quarks) transform as doublets with respect to the  $SU(2)$  factor whereas the right-handed components are singlets. The charge with respect to the  $U(1)$  part, the hypercharge  $Y$ , is determined by the Gell-Mann-Nishijima formula

$$Q = T_3 + Y, \quad (357)$$

where  $Q$  is the electric charge of the corresponding particle and  $T_3$  is the eigenvalue with respect to the third generator of the  $SU(2)$  group in the corresponding representation:  $T_3 = \frac{1}{2}\sigma^3$  for the doublets and  $T_3 = 0$  for the singlets. For the first family of quarks ( $u$ ,  $d$ ) and leptons ( $e$ ,  $\nu_e$ ) we have the following field content

$$\begin{aligned} \text{quarks:} & \quad \left( \begin{array}{c} u^\alpha \\ d^\alpha \end{array} \right)_{L, \frac{1}{6}} & \quad u_{R, \frac{2}{3}}^\alpha & \quad d_{R, \frac{2}{3}}^\alpha \\ \text{leptons:} & \quad \left( \begin{array}{c} \nu_e \\ e \end{array} \right)_{L, -\frac{1}{2}} & \quad e_{R, -1} \end{aligned} \quad (358)$$

where  $\alpha = 1, 2, 3$  labels the color quantum number and the subscript indicates the value of the weak hypercharge  $Y$ . Denoting the representations of  $SU(3) \times SU(2) \times U(1)$  by  $(n_c, n_w)_Y$ , with  $n_c$  and  $n_w$  the representations of  $SU(3)$  and  $SU(2)$  respectively and  $Y$  the hypercharge, the matter content of the standard model consists of a three family replication of the representations:

$$\begin{aligned} \text{left-handed fermions:} & \quad (3, 2)_{\frac{1}{6}}^L & \quad (1, 2)_{-\frac{1}{2}}^L \\ \text{right-handed fermions:} & \quad (3, 1)_{\frac{2}{3}}^R & \quad (3, 1)_{-\frac{1}{3}}^R & \quad (1, 1)_{-1}^R. \end{aligned} \quad (359)$$

In computing the triangle diagram we have 10 possibilities depending on which factor of the gauge group

$SU(3) \times SU(2) \times U(1)$  couples to each vertex:

$$\begin{array}{lll}
 SU(3)^3 & SU(2)^3 & U(1)^3 \\
 SU(3)^2 SU(2) & SU(2)^2 U(1) & \\
 SU(3)^2 U(1) & SU(2) U(1)^2 & \\
 SU(3) SU(2)^2 & & \\
 SU(3) SU(2) U(1) & & \\
 SU(3) U(1)^2 & & 
 \end{array}$$

It is easy to check that some of them do not give rise to anomalies. For example the anomaly for the  $SU(3)^3$  case cancels because left and right-handed quarks transform in the same representation. In the case of  $SU(2)^3$  the cancellation happens term by term because of the Pauli matrices identity  $\sigma^a \sigma^b = \delta^{ab} + i \varepsilon^{abc} \sigma^c$  that leads to

$$\text{tr} \left[ \sigma^a \{ \sigma^b, \sigma^c \} \right] = 2 (\text{tr} \sigma^a) \delta^{bc} = 0. \quad (360)$$

However the hardest anomaly cancellation condition to satisfy is the one with three  $U(1)$ 's. In this case the absence of anomalies within a single family is guaranteed by the nontrivial identity

$$\begin{aligned}
 \sum_{\text{left}} Y_+^3 - \sum_{\text{right}} Y_-^3 &= 3 \times 2 \times \left( \frac{1}{6} \right)^3 + 2 \times \left( -\frac{1}{2} \right)^3 - 3 \times \left( \frac{2}{3} \right)^3 - 3 \times \left( -\frac{1}{3} \right)^3 - (-1)^3 \\
 &= \left( -\frac{3}{4} \right) + \left( \frac{3}{4} \right) = 0.
 \end{aligned} \quad (361)$$

It is remarkable that the anomaly exactly cancels between leptons and quarks. Notice that this result holds even if a right-handed sterile neutrino is added since such a particle is a singlet under the whole standard model gauge group and therefore does not contribute to the triangle diagram. Therefore we see how the matter content of the standard model conspires to yield a consistent quantum field theory.

In all our discussion of anomalies we only considered the computation of one-loop diagrams. It may happen that higher loop orders impose additional conditions. Fortunately this is not so: the Adler-Bardeen theorem [35] guarantees that the axial anomaly only receives contributions from one loop diagrams. Therefore, once anomalies are canceled (if possible) at one loop we know that there will be no new conditions coming from higher-loop diagrams in perturbation theory.

The Adler-Bardeen theorem, however, only applies in perturbation theory. It is nonetheless possible that nonperturbative effects can result in the quantum violation of a gauge symmetry. This is precisely the case pointed out by Witten [36] with respect to the  $SU(2)$  gauge symmetry of the standard model. In this case the problem lies in the nontrivial topology of the gauge group  $SU(2)$ . The invariance of the theory with respect to gauge transformations which are not in the connected component of the identity makes all correlation functions equal to zero. Only when the number of left-handed  $SU(2)$  fermion doublets is even gauge invariance allows for a nontrivial theory. It is again remarkable that the family structure of the standard model makes this anomaly to cancel

$$3 \times \begin{pmatrix} u \\ d \end{pmatrix}_L + 1 \times \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L = 4 \text{ SU(2)-doublets}, \quad (362)$$

where the factor of 3 comes from the number of colors.

## 8 Renormalization

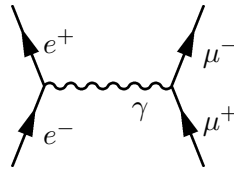
### 8.1 Removing infinities

From its very early stages, quantum field theory was faced with infinities. They emerged in the calculation of most physical quantities, such as the correction to the charge of the electron due to the interactions with the radiation field. The way these divergences were handled in the 1940s, starting with Kramers, was physically very much in the spirit of the Quantum Theory emphasis in observable quantities: since the observed magnitude of physical quantities (such as the charge of the electron) is finite, this number should arise from the addition of a “bare” (unobservable) value and the quantum corrections. The fact that both of these quantities were divergent was not a problem physically, since only its finite sum was an observable quantity. To make things mathematically sound, the handling of infinities requires the introduction of some regularization procedure which cuts the divergent integrals off at some momentum scale  $\Lambda$ . Morally speaking, the physical value of an observable  $\mathcal{O}_{\text{physical}}$  is given by

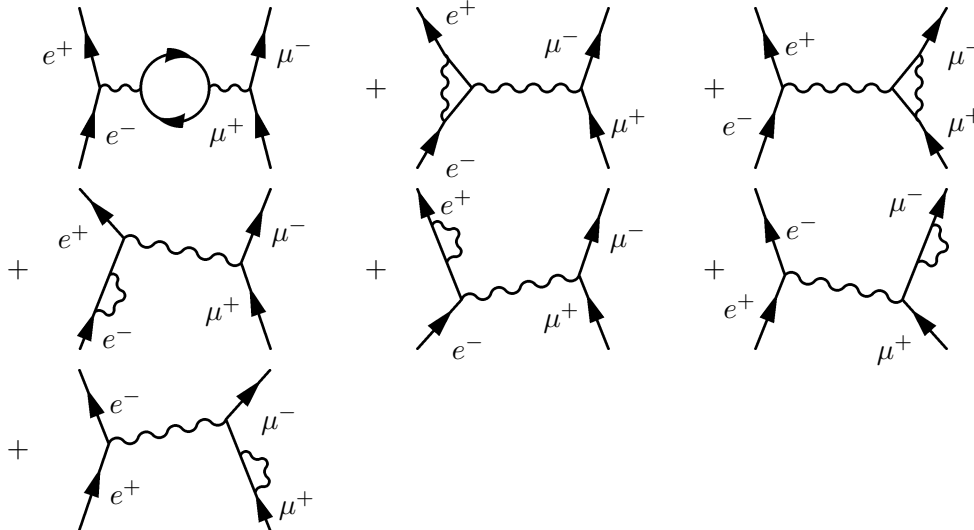
$$\mathcal{O}_{\text{physical}} = \lim_{\Lambda \rightarrow \infty} [\mathcal{O}(\Lambda)_{\text{bare}} + \Delta\mathcal{O}(\Lambda)_{\hbar}], \quad (363)$$

where  $\Delta\mathcal{O}(\Lambda)_{\hbar}$  represents the regularized quantum corrections.

To make this qualitative discussion more precise we compute the corrections to the electric charge in Quantum Electrodynamics. We consider the process of annihilation of an electron-positron pair to create a muon-antimuon pair  $e^-e^+ \rightarrow \mu^+\mu^-$ . To lowest order in the electric charge  $e$  the only diagram contributing is



However, the corrections at order  $e^4$  to this result requires the calculation of seven more diagrams



In order to compute the renormalization of the charge we consider the first diagram which takes into account the first correction to the propagator of the virtual photon interchanged between the pairs

due to vacuum polarization. We begin by evaluating

$$\begin{array}{c} \text{---} \circlearrowleft \text{---} \end{array} = \frac{-i\eta^{\mu\alpha}}{q^2 + i\epsilon} \left[ \begin{array}{c} \alpha \circlearrowleft \beta \end{array} \right] \frac{-i\eta^{\beta\nu}}{q^2 + i\epsilon}, \quad (364)$$

where the diagram between brackets is given by

$$\begin{array}{c} \alpha \circlearrowleft \beta \end{array} \equiv \Pi^{\alpha\beta}(q) = i^2(-ie)^2(-1) \int \frac{d^4k}{(2\pi)^4} \frac{\text{Tr}[(\not{k} + m_e)\gamma^\alpha(\not{k} + \not{q} + m_e)\gamma^\beta]}{[k^2 - m_e^2 + i\epsilon][(k+q)^2 - m_e^2 + i\epsilon]}. \quad (365)$$

Physically this diagram includes the correction to the propagator due to the polarization of the vacuum, i.e. the creation of virtual electron-positron pairs by the propagating photon. The momentum  $q$  is the total momentum of the electron-positron pair in the intermediate channel.

It is instructive to look at this diagram from the point of view of perturbation theory in nonrelativistic Quantum Mechanics. In each vertex the interaction consists of the annihilation (resp. creation) of a photon and the creation (resp. annihilation) of an electron-positron pair. This can be implemented by the interaction Hamiltonian

$$H_{\text{int}} = e \int d^3x \bar{\psi} \gamma^\mu \psi A_\mu. \quad (366)$$

All fields inside the integral can be expressed in terms of the corresponding creation-annihilation operators for photons, electrons and positrons. In Quantum Mechanics, the change in the wave function at first order in the perturbation  $H_{\text{int}}$  is given by

$$|\gamma, \text{in}\rangle = |\gamma, \text{in}\rangle_0 + \sum_n \frac{\langle n | H_{\text{int}} | \gamma, \text{in}\rangle_0}{E_{\text{in}} - E_n} |n\rangle \quad (367)$$

and similarly for  $|\gamma, \text{out}\rangle$ , where we have denoted symbolically by  $|n\rangle$  all the possible states of the electron-positron pair. Since these states are orthogonal to  $|\gamma, \text{in}\rangle_0$ ,  $|\gamma, \text{out}\rangle_0$ , we find to order  $e^2$

$$\langle \gamma, \text{in} | \gamma', \text{out} \rangle = {}_0\langle \gamma, \text{in} | \gamma', \text{out} \rangle_0 + \sum_n \frac{{}_0\langle \gamma, \text{in} | H_{\text{int}} | n \rangle \langle n | H_{\text{int}} | \gamma', \text{out} \rangle_0}{(E_{\text{in}} - E_n)(E_{\text{out}} - E_n)} + \mathcal{O}(e^4). \quad (368)$$

Hence, we see that the diagram of Eq. (364) really corresponds to the order- $e^2$  correction to the photon propagator  $\langle \gamma, \text{in} | \gamma', \text{out} \rangle$

$$\begin{array}{c} \text{---} \gamma \text{---} \gamma' \text{---} \\ \text{---} \gamma \text{---} \circlearrowleft \text{---} \gamma' \text{---} \end{array} \begin{array}{l} \longrightarrow {}_0\langle \gamma, \text{in} | \gamma', \text{out} \rangle_0 \\ \longrightarrow \sum_n \frac{\langle \gamma, \text{in} | H_{\text{int}} | n \rangle \langle n | H_{\text{int}} | \gamma', \text{out} \rangle}{(E_{\text{in}} - E_n)(E_{\text{out}} - E_n)}. \end{array} \quad (369)$$



Once we understood the physical meaning of the Feynman diagram to be computed we proceed to its evaluation. In principle there is no problem in computing the integral in Eq. (364) for nonzero values of the electron mass. However since here we are going to be mostly interested in seeing how the divergence of the integral results in a scale-dependent renormalization of the electric charge, we will set  $m_e = 0$ . This is something safe to do, since in the case of this diagram we are not inducing new infrared divergences in taking the electron as massless. Implementing gauge invariance and using standard techniques in the computation of Feynman diagrams (see references [1]- [11]) the polarization tensor  $\Pi_{\mu\nu}(q)$  defined in Eq. (365) can be written as

$$\Pi_{\mu\nu}(q) = (q^2 \eta_{\mu\nu} - q_\mu q_\nu) \Pi(q^2) \quad (370)$$

with

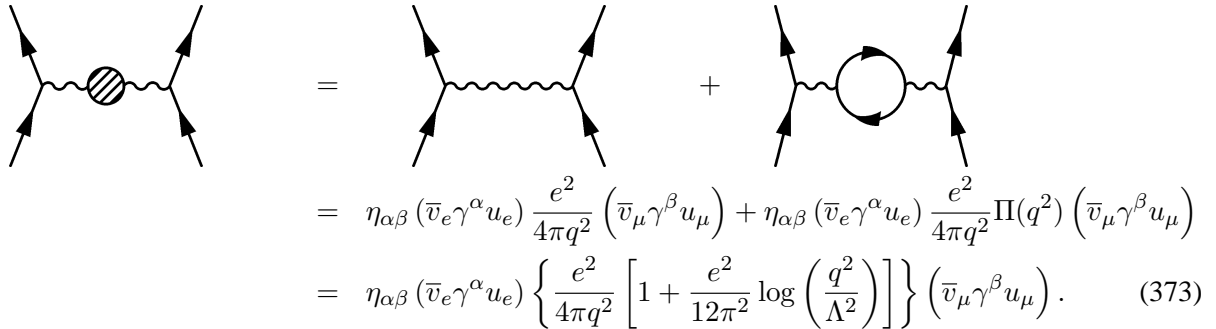
$$\Pi(q) = 8e^2 \int_0^1 dx \int \frac{d^4 k}{(2\pi)^4} \frac{x(1-x)}{[k^2 - m^2 + x(1-x)q^2 + i\epsilon]^2} \quad (371)$$

To handle this divergent integral we have to figure out some procedure to render it finite. This can be done in several ways, but here we choose to cut the integrals off at a high energy scale  $\Lambda$ , where new physics might be at work,  $|p| < \Lambda$ . This gives the result

$$\Pi(q^2) \simeq \frac{e^2}{12\pi^2} \log\left(\frac{q^2}{\Lambda^2}\right) + \text{finite terms.} \quad (372)$$

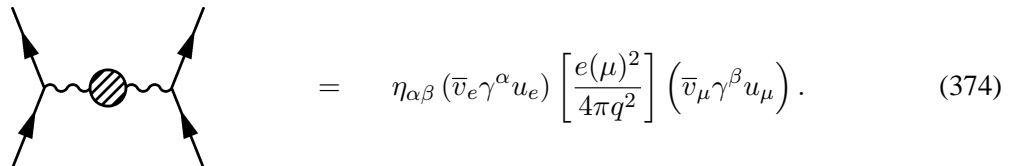
If we would send the cutoff to infinity  $\Lambda \rightarrow \infty$  the divergence blows up and something has to be done about it.

If we want to make sense out of this, we have to go back to the physical question that led us to compute Eq. (364). Our primordial motivation was to compute the corrections to the annihilation of two electrons into two muons. Including the correction to the propagator of the virtual photon we have



$$\begin{aligned} &= \eta_{\alpha\beta} (\bar{v}_e \gamma^\alpha u_e) \frac{e^2}{4\pi q^2} (\bar{v}_\mu \gamma^\beta u_\mu) + \eta_{\alpha\beta} (\bar{v}_e \gamma^\alpha u_e) \frac{e^2}{4\pi q^2} \Pi(q^2) (\bar{v}_\mu \gamma^\beta u_\mu) \\ &= \eta_{\alpha\beta} (\bar{v}_e \gamma^\alpha u_e) \left\{ \frac{e^2}{4\pi q^2} \left[ 1 + \frac{e^2}{12\pi^2} \log\left(\frac{q^2}{\Lambda^2}\right) \right] \right\} (\bar{v}_\mu \gamma^\beta u_\mu). \end{aligned} \quad (373)$$

Now let us imagine that we are performing a  $e^- e^+ \rightarrow \mu^- \mu^+$  with a center of mass energy  $\mu$ . From the previous result we can identify the effective charge of the particles at this energy scale  $e(\mu)$  as



$$= \eta_{\alpha\beta} (\bar{v}_e \gamma^\alpha u_e) \left[ \frac{e(\mu)^2}{4\pi q^2} \right] (\bar{v}_\mu \gamma^\beta u_\mu). \quad (374)$$

This charge,  $e(\mu)$ , is the quantity that is physically measurable in our experiment. Now we can make sense of the formally divergent result (373) by assuming that the charge appearing in the classical Lagrangian of QED is just a “bare” value that depends on the scale  $\Lambda$  at which we cut off the theory,  $e \equiv e(\Lambda)_{\text{bare}}$ . In order to reconcile (373) with the physical results (374) we must assume that the dependence of the bare (unobservable) charge  $e(\Lambda)_{\text{bare}}$  on the cutoff  $\Lambda$  is determined by the identity

$$e(\mu)^2 = e(\Lambda)_{\text{bare}}^2 \left[ 1 + \frac{e(\Lambda)_{\text{bare}}^2}{12\pi^2} \log\left(\frac{\mu^2}{\Lambda^2}\right) \right]. \quad (375)$$

If we still insist in removing the cutoff,  $\Lambda \rightarrow \infty$  we have to send the bare charge to zero  $e(\Lambda)_{\text{bare}} \rightarrow 0$  in such a way that the effective coupling has the finite value given by the experiment at the energy scale  $\mu$ . It is not a problem, however, that the bare charge is small for large values of the cutoff, since the only measurable quantity is the effective charge that remains finite. Therefore all observable quantities should be expressed in perturbation theory as a power series in the physical coupling  $e(\mu)^2$  and not in the unphysical bare coupling  $e(\Lambda)_{\text{bare}}$ .

## 8.2 The beta-function and asymptotic freedom

We can look at the previous discussion, in particular Eq. (375), from a different point of view. In order to remove the ambiguities associated with infinities we have been forced to introduce a dependence of the coupling constant on the energy scale at which a process takes place. From the expression of the physical coupling in terms of the bare charge (375) we can actually eliminate the cutoff  $\Lambda$ , whose value after all should not affect the value of physical quantities. Taking into account that we are working in perturbation theory in  $e(\mu)^2$ , we can express the bare charge  $e(\Lambda)_{\text{bare}}^2$  in terms of  $e(\mu)^2$  as

$$e(\Lambda)^2 = e(\mu)^2 \left[ 1 + \frac{e(\mu)^2}{12\pi^2} \log \left( \frac{\mu^2}{\Lambda^2} \right) \right] + \mathcal{O}[e(\mu)^6]. \quad (376)$$

This expression allow us to eliminate all dependence in the cutoff in the expression of the effective charge at a scale  $\mu$  by replacing  $e(\Lambda)_{\text{bare}}$  in Eq. (375) by the one computed using (376) at a given reference energy scale  $\mu_0$

$$e(\mu)^2 = e(\mu_0)^2 \left[ 1 + \frac{e(\mu_0)^2}{12\pi^2} \log \left( \frac{\mu^2}{\mu_0^2} \right) \right]. \quad (377)$$

From this equation we can compute, at this order in perturbation theory, the effective value of the coupling constant at an energy  $\mu$ , once we know its value at some reference energy scale  $\mu_0$ . In the case of the electron charge we can use as a reference Thompson's scattering at energies of the order of the electron mass  $m_e \simeq 0.5$  MeV, at where the value of the electron charge is given by the well known value

$$e(m_e)^2 \simeq \frac{1}{137}. \quad (378)$$

With this we can compute  $e(\mu)^2$  at any other energy scale applying Eq. (377), for example at the electron mass  $\mu = m_e \simeq 0.5$  MeV. However, in computing the electromagnetic coupling constant at any other scale we must take into account the fact that other charged particles can run in the loop in Eq. (373). Suppose, for example, that we want to calculate the fine structure constant at the mass of the  $Z^0$ -boson  $\mu = M_Z \equiv 92$  GeV. Then we should include in Eq. (377) the effect of other fermionic standard model fields with masses below  $M_Z$ . Doing this, we find<sup>18</sup>

$$e(M_Z)^2 = e(m_e)^2 \left[ 1 + \frac{e(m_e)^2}{12\pi^2} \left( \sum_i q_i^2 \right) \log \left( \frac{M_Z^2}{m_e^2} \right) \right], \quad (379)$$

where  $q_i$  is the charge in units of the electron charge of the  $i$ -th fermionic species running in the loop and we sum over all fermions with masses below the mass of the  $Z^0$  boson. This expression shows how the electromagnetic coupling grows with energy. However, in order to compare with the experimental value of  $e(M_Z)^2$  it is not enough with including the effect of fermionic fields, since also the  $W^\pm$  bosons

<sup>18</sup>In the first version of these notes the argument used to show the growing of the electromagnetic coupling constant could have led to confusion to some readers. To avoid this potential problem we include in the equation for the running coupling  $e(\mu)^2$  the contribution of all fermions with masses below  $M_Z$ . We thank Lubos Motl for bringing this issue to our attention.

can run in the loop ( $M_W < M_Z$ ). Taking this into account, as well as threshold effects, the value of the electron charge at the scale  $M_Z$  is found to be [37]

$$e(M_Z)^2 \simeq \frac{1}{128.9} . \quad (380)$$

This growing of the effective fine structure constant with energy can be understood heuristically by remembering that the effect of the polarization of the vacuum shown in the diagram of Eq. (364) amounts to the creation of a plethora of electron-positron pairs around the location of the charge. These virtual pairs behave as dipoles that, as in a dielectric medium, tend to screen this charge and decreasing its value at long distances (i.e. lower energies).

The variation of the coupling constant with energy is usually encoded in quantum field theory in the *beta function* defined by

$$\beta(g) = \mu \frac{dg}{d\mu} . \quad (381)$$

In the case of QED the beta function can be computed from Eq. (377) with the result

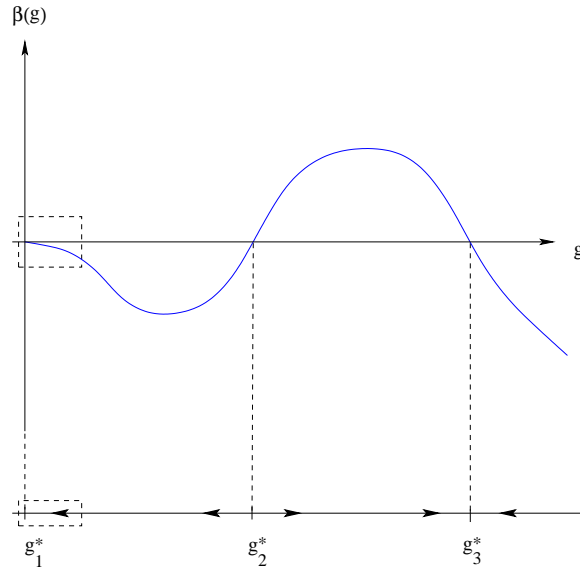
$$\beta(e)_{\text{QED}} = \frac{e^3}{12\pi^2} . \quad (382)$$

The fact that the coefficient of the leading term in the beta-function is positive  $\beta_0 \equiv \frac{1}{6\pi} > 0$  gives us the overall behavior of the coupling as we change the scale. Eq. (382) means that, if we start at an energy where the electric coupling is small enough for our perturbative treatment to be valid, the effective charge grows with the energy scale. This growing of the effective coupling constant with energy means that QED is infrared safe, since the perturbative approximation gives better and better results as we go to lower energies. Actually, because the electron is the lighter electrically charged particle and has a finite nonvanishing mass the running of the fine structure constant stops at the scale  $m_e$  in the well-known value  $\frac{1}{137}$ . Would other charged fermions with masses below  $m_e$  be present in Nature, the effective value of the fine structure constant in the interaction between these particles would run further to lower values at energies below the electron mass.

On the other hand if we increase the energy scale  $e(\mu)^2$  grows until at some scale the coupling is of order one and the perturbative approximation breaks down. In QED this is known as the problem of the Landau pole but in fact it does not pose any serious threat to the reliability of QED perturbation theory: a simple calculation shows that the energy scale at which the theory would become strongly coupled is  $\Lambda_{\text{Landau}} \simeq 10^{277}$  GeV. However, we know that QED does not live that long! At much lower scales we expect electromagnetism to be unified with other interactions, and even if this is not the case we will enter the uncharted territory of quantum gravity at energies of the order of  $10^{19}$  GeV.

So much for QED. The next question that one may ask at this stage is whether it is possible to find quantum field theories with a behavior opposite to that of QED, i.e. such that they become weakly coupled at high energies. This is not a purely academic question. In the late 1960s a series of deep-inelastic scattering experiments carried out at SLAC showed that the quarks behave essentially as free particles inside hadrons. The apparent problem was that no theory was known at that time that would become free at very short distances: the example set by QED seem to be followed by all the theories that were studied. This posed a very serious problem for quantum field theory as a way to describe subnuclear physics, since it seemed that its predictive power was restricted to electrodynamics but failed miserably when applied to describe strong interactions.

Nevertheless, this critical time for quantum field theory turned out to be its finest hour. In 1973 David Gross and Frank Wilczek [38] and David Politzer [39] showed that nonabelian gauge theories can actually display the required behavior. For the QCD Lagrangian in Eq. (331) the beta function is given



**Fig. 14:** Beta function for a hypothetical theory with three fixed points  $g_1^*$ ,  $g_2^*$  and  $g_3^*$ . A perturbative analysis would capture only the regions shown in the boxes.

by<sup>19</sup>

$$\beta(g) = -\frac{g^3}{16\pi^2} \left[ \frac{11}{3}N_c - \frac{2}{3}N_f \right]. \quad (383)$$

In particular, for real QCD ( $N_C = 3$ ,  $N_f = 6$ ) we have that  $\beta(g) = -\frac{7g^3}{16\pi^2} < 0$ . This means that for a theory that is weakly coupled at an energy scale  $\mu_0$  the coupling constant decreases as the energy increases  $\mu \rightarrow \infty$ . This explains the apparent freedom of quarks inside the hadrons: when the quarks are very close together their effective color charge tends to zero. This phenomenon is called *asymptotic freedom*.

Asymptotically free theories display a behavior that is opposite to that found above in QED. At high energies their coupling constant approaches zero whereas at low energies they become strongly coupled (infrared slavery). These features are at the heart of the success of QCD as a theory of strong interactions, since this is exactly the type of behavior found in quarks: they are quasi-free particles inside the hadrons but the interaction potential between them increases at large distances.

Although asymptotically free theories can be handled in the ultraviolet, they become extremely complicated in the infrared. In the case of QCD it is still to be understood (at least analytically) how the theory confines color charges and generates the spectrum of hadrons, as well as the breaking of the chiral symmetry (345).

In general, the ultraviolet and infrared properties of a theory are controlled by the fixed points of the beta function, i.e. those values of the coupling constant  $g$  for which it vanishes

$$\beta(g^*) = 0. \quad (384)$$

Using perturbation theory we have seen that for both QED and QCD one of such fixed points occurs at zero coupling,  $g^* = 0$ . However, our analysis also showed that the two theories present radically different behavior at high and low energies. From the point of view of the beta function, the difference lies in the energy regime at which the coupling constant approaches its critical value. This is in fact governed by the sign of the beta function around the critical coupling.

<sup>19</sup>The expression of the beta function of QCD was also known to 't Hooft [40]. There are even earlier computations in the Russian literature [41].

We have seen above that when the beta function is negative close to the fixed point (the case of QCD) the coupling tends to its critical value,  $g^* = 0$ , as the energy is increased. This means that the critical point is *ultraviolet stable*, i.e. it is an attractor as we evolve towards higher energies. If, on the contrary, the beta function is positive (as it happens in QED) the coupling constant approaches the critical value as the energy decreases. This is the case of an *infrared stable* fixed point.

This analysis that we have motivated with the examples of QED and QCD is completely general and can be carried out for any quantum field theory. In Fig. 14 we have represented the beta function for a hypothetical theory with three fixed points located at couplings  $g_1^*$ ,  $g_2^*$  and  $g_3^*$ . The arrows in the line below the plot represent the evolution of the coupling constant as the energy increases. From the analysis presented above we see that  $g_1^* = 0$  and  $g_3^*$  are ultraviolet stable fixed points, while the fixed point  $g_2^*$  is infrared stable.

In order to understand the high and low energy behavior of a quantum field theory it is then crucial to know the structure of the beta functions associated with its couplings. This can be a very difficult task, since perturbation theory only allows the study of the theory around “trivial” fixed points, i.e. those that occur at zero coupling like the case of  $g_1^*$  in Fig. 14. On the other hand, any “nontrivial” fixed point occurring in a theory (like  $g_2^*$  and  $g_3^*$ ) cannot be captured in perturbation theory and requires a full nonperturbative analysis.

The moral to be learned from our discussion above is that dealing with the ultraviolet divergences in a quantum field theory has the consequence, among others, of introducing an energy dependence in the measured value of the coupling constants of the theory (for example the electric charge in QED). This happens even in the case of renormalizable theories without mass terms. These theories are scale invariant at the classical level because the action does not contain any dimensionful parameter. In this case the running of the coupling constants can be seen as resulting from a quantum breaking of classical scale invariance: different energy scales in the theory are distinguished by different values of the coupling constants. Remembering what we learned in Section 7, we conclude that classical scale invariance is an anomalous symmetry. One heuristic way to see how the conformal anomaly comes about is to notice that the regularization of an otherwise scale invariant field theory requires the introduction of an energy scale (e.g. a cutoff). This breaking of scale invariance cannot be restored after renormalization.

Nevertheless, scale invariance is not lost forever in the quantum theory. It is recovered at the fixed points of the beta function where, by definition, the coupling does not run. To understand how this happens we go back to a scale invariant classical field theory whose field  $\phi(x)$  transform under coordinate rescalings as

$$x^\mu \longrightarrow \lambda x^\mu, \quad \phi(x) \longrightarrow \lambda^{-\Delta} \phi(\lambda^{-1}x), \quad (385)$$

where  $\Delta$  is called the canonical scaling dimension of the field. An example of such a theory is a massless  $\phi^4$  theory in four dimensions

$$\mathcal{L} = \frac{1}{2} \partial_\mu \phi \partial^\mu \phi - \frac{g}{4!} \phi^4, \quad (386)$$

where the scalar field has canonical scaling dimension  $\Delta = 1$ . The Lagrangian density transforms as

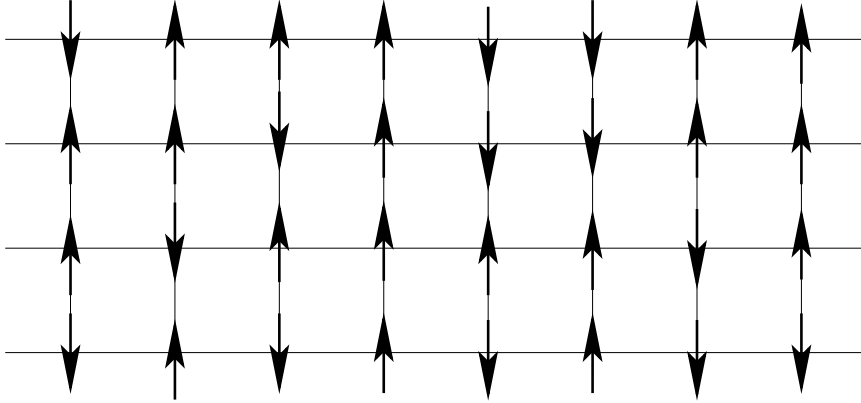
$$\mathcal{L} \longrightarrow \lambda^{-4} \mathcal{L}[\phi] \quad (387)$$

and the classical action remains invariant<sup>20</sup>.

If scale invariance is preserved under quantization, the Green's functions transform as

$$\langle \Omega | T[\phi'(x_1) \dots \phi'(x_n)] | \Omega \rangle = \lambda^{n\Delta} \langle \Omega | T[\phi(\lambda^{-1}x_1) \dots \phi(\lambda^{-1}x_n)] | \Omega \rangle. \quad (388)$$

<sup>20</sup>In a  $D$ -dimensional theory the canonical scaling dimensions of the fields coincide with its engineering dimension:  $\Delta = \frac{D-2}{2}$  for bosonic fields and  $\Delta = \frac{D-1}{2}$  for fermionic ones. For a Lagrangian with no dimensionful parameters classical scale invariance follows then from dimensional analysis.



**Fig. 15:** Systems of spins in a two-dimensional square lattice.

This is precisely what happens in a free theory. In an interacting theory the running of the coupling constant destroys classical scale invariance at the quantum level. Despite of this, at the fixed points of the beta function the Green's functions transform again according to (388) where  $\Delta$  is replaced by

$$\Delta_{\text{anom}} = \Delta + \gamma^*. \quad (389)$$

The canonical scaling dimension of the fields are corrected by  $\gamma^*$ , which is called the anomalous dimension. They carry the dynamical information about the high-energy behavior of the theory.

### 8.3 The renormalization group

In spite of its successes, the renormalization procedure presented above can be seen as some kind of prescription or recipe to get rid of the divergences in an ordered way. This discomfort about renormalization was expressed in occasions by comparing it with “sweeping the infinities under the rug”. However thanks to Ken Wilson to a large extent [42] the process of renormalization is now understood in a very profound way as a procedure to incorporate the effects of physics at high energies by modifying the value of the parameters that appear in the Lagrangian.

**Statistical mechanics.** Wilson's ideas are both simple and profound and consist in thinking about quantum field theory as the analog of a thermodynamical description of a statistical system. To be more precise, let us consider an Ising spin system in a two-dimensional square lattice as the one depicted in Fig 15. In terms of the spin variables  $s_i = \pm \frac{1}{2}$ , where  $i$  labels the lattice site, the Hamiltonian of the system is given by

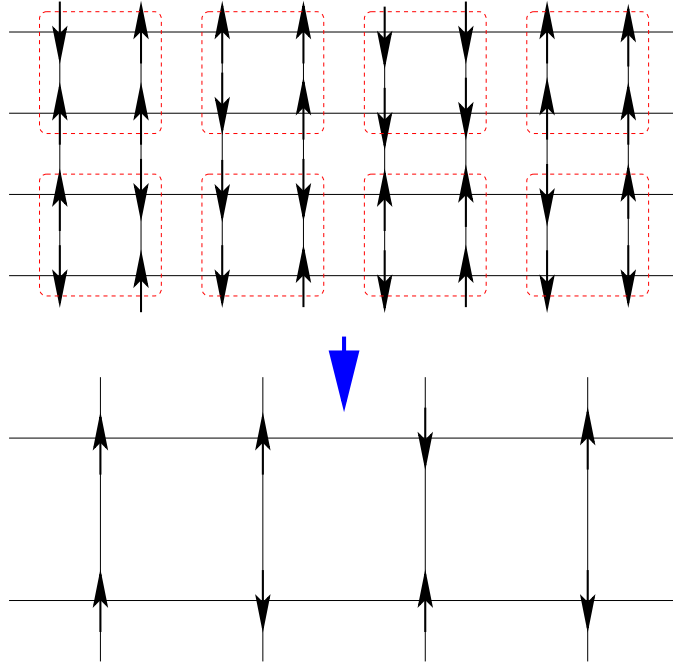
$$H = -J \sum_{\langle i,j \rangle} s_i s_j, \quad (390)$$

where  $\langle i,j \rangle$  indicates that the sum extends over nearest neighbors and  $J$  is the coupling constant between neighboring spins (here we consider that there is no external magnetic field). The starting point to study the statistical mechanics of this system is the partition function defined as

$$\mathcal{Z} = \sum_{\{s_i\}} e^{-\beta H}, \quad (391)$$

where the sum is over all possible configurations of the spins and  $\beta = \frac{1}{T}$  is the inverse temperature. For  $J > 0$  the Ising model presents spontaneous magnetization below a critical temperature  $T_c$ , in any dimension higher than one. Away from this temperature correlations between spins decay exponentially at large distances

$$\langle s_i s_j \rangle \sim e^{-\frac{|x_{ij}|}{\xi}}, \quad (392)$$



**Fig. 16:** Decimation of the spin lattice. Each block in the upper lattice is replaced by an effective spin computed according to the rule (394). Notice also that the size of the lattice spacing is doubled in the process.

with  $|x_{ij}|$  the distance between the spins located in the  $i$ -th and  $j$ -th sites of the lattice. This expression serves as a definition of the correlation length  $\xi$  which sets the characteristic length scale at which spins can influence each other by their interaction through their nearest neighbors.

Suppose now that we are interested in a macroscopic description of this spin system. We can capture the relevant physics by integrating out somehow the physics at short scales. A way in which this can be done was proposed by Leo Kadanoff [43] and consists in dividing our spin system in spin-blocks like the ones showed in Fig 16. Now we can construct another spin system where each spin-block of the original lattice is replaced by an effective spin calculated according to some rule from the spins contained in each block  $B_a$

$$\{s_i : i \in B_a\} \longrightarrow s_a^{(1)}. \quad (393)$$

For example we can define the effective spin associated with the block  $B_a$  by taking the majority rule with an additional prescription in case of a draw

$$s_a^{(1)} = \frac{1}{2} \text{sgn} \left( \sum_{i \in B_a} s_i \right), \quad (394)$$

where we have used the sign function,  $\text{sgn}(x) \equiv \frac{x}{|x|}$ , with the additional definition  $\text{sgn}(0) = 1$ . This procedure is called decimation and leads to a new spin system with a doubled lattice space.

The idea now is to rewrite the partition function (391) only in terms of the new effective spins  $s_a^{(1)}$ . Then we start by splitting the sum over spin configurations into two nested sums, one over the spin blocks and a second one over the spins within each block

$$\mathcal{Z} = \sum_{\{\vec{s}\}} e^{-\beta H[\vec{s}]} = \sum_{\{\vec{s}^{(1)}\}} \sum_{\{\vec{s} \in B_a\}} \delta \left[ s_a^{(1)} - \text{sgn} \left( \sum_{i \in B_a} s_i \right) \right] e^{-\beta H[\vec{s}]} \quad (395)$$

The interesting point now is that the sum over spins inside each block can be written as the exponential of a new effective Hamiltonian depending only on the effective spins,  $H^{(1)}[s_a^{(1)}]$

$$\sum_{\{s \in B_a\}} \delta \left[ s_a^{(1)} - \text{sign} \left( \sum_{i \in B_a} s_i \right) \right] e^{-\beta H[s_i]} = e^{-\beta H^{(1)}[s_a^{(1)}]}. \quad (396)$$

The new Hamiltonian is of course more complicated

$$H^{(1)} = -J^{(1)} \sum_{\langle i,j \rangle} s_i^{(1)} s_j^{(1)} + \dots \quad (397)$$

where the dots stand for other interaction terms between the effective block spins. This new terms appear because in the process of integrating out short distance physics we induce interactions between the new effective degrees of freedom. For example the interaction between the spin block variables  $s_i^{(1)}$  will in general not be restricted to nearest neighbors in the new lattice. The important point is that we have managed to rewrite the partition function solely in terms of this new (renormalized) spin variables  $s^{(1)}$  interacting through a new Hamiltonian  $H^{(1)}$

$$\mathcal{Z} = \sum_{\{s^{(1)}\}} e^{-\beta H^{(1)}[s_a^{(1)}]}. \quad (398)$$

Let us now think about the space of all possible Hamiltonians for our statistical system including all kinds of possible couplings between the individual spins compatible with the symmetries of the system. If denote by  $\mathcal{R}$  the decimation operation, our previous analysis shows that  $\mathcal{R}$  defines a map in this space of Hamiltonians

$$\mathcal{R} : H \rightarrow H^{(1)}. \quad (399)$$

At the same time the operation  $\mathcal{R}$  replaces a lattice with spacing  $a$  by another one with double spacing  $2a$ . As a consequence the correlation length in the new lattice measured in units of the lattice spacing is divided by two,  $\mathcal{R} : \xi \rightarrow \frac{\xi}{2}$ .

Now we can iterate the operation  $\mathcal{R}$  an indefinite number of times. Eventually we might reach a Hamiltonian  $H_*$  that is not further modified by the operation  $\mathcal{R}$

$$H \xrightarrow{\mathcal{R}} H^{(1)} \xrightarrow{\mathcal{R}} H^{(2)} \xrightarrow{\mathcal{R}} \dots \xrightarrow{\mathcal{R}} H_*. \quad (400)$$

The fixed point Hamiltonian  $H_*$  is *scale invariant* because it does not change as  $\mathcal{R}$  is performed. Notice that because of this invariance the correlation length of the system at the fixed point do not change under  $\mathcal{R}$ . This fact is compatible with the transformation  $\xi \rightarrow \frac{\xi}{2}$  only if  $\xi = 0$  or  $\xi = \infty$ . Here we will focus in the case of nontrivial fixed points with infinite correlation length.

The space of Hamiltonians can be parametrized by specifying the values of the coupling constants associated with all possible interaction terms between individual spins of the lattice. If we denote by  $\mathcal{O}_a[s_i]$  these (possibly infinite) interaction terms, the most general Hamiltonian for the spin system under study can be written as

$$H[s_i] = \sum_{a=1}^{\infty} \lambda_a \mathcal{O}_a[s_i], \quad (401)$$

where  $\lambda_a \in \mathbb{R}$  are the coupling constants for the corresponding operators. These constants can be thought of as coordinates in the space of all Hamiltonians. Therefore the operation  $\mathcal{R}$  defines a transformation in the set of coupling constants

$$\mathcal{R} : \lambda_a \longrightarrow \lambda_a^{(1)}. \quad (402)$$



For example, in our case we started with a Hamiltonian in which only one of the coupling constants is different from zero (say  $\lambda_1 = -J$ ). As a result of the decimation  $\lambda_1 \equiv -J \rightarrow -J^{(1)}$  while some of the originally vanishing coupling constants will take a nonzero value. Of course, for the fixed point Hamiltonian the coupling constants do not change under the scale transformation  $\mathcal{R}$ .

Physically the transformation  $\mathcal{R}$  integrates out short distance physics. The consequence for physics at long distances is that we have to replace our Hamiltonian by a new one with different values for the coupling constants. That is, our ignorance of the details of the physics going on at short distances result in a *renormalization* of the coupling constants of the Hamiltonian that describes the long range physical processes. It is important to stress that although  $\mathcal{R}$  is sometimes called a renormalization group transformation in fact this is a misnomer. Transformations between Hamiltonians defined by  $\mathcal{R}$  do not form a group: since these transformations proceed by integrating out degrees of freedom at short scales they cannot be inverted.

In statistical mechanics fixed points under renormalization group transformations with  $\xi = \infty$  are associated with phase transitions. From our previous discussion we can conclude that the space of Hamiltonians is divided in regions corresponding to the basins of attraction of the different fixed points. We can ask ourselves now about the stability of those fixed points. Suppose we have a statistical system described by a fixed-point Hamiltonian  $H_*$  and we perturb it by changing the coupling constant associated with an interaction term  $\mathcal{O}$ . This is equivalent to replace  $H_*$  by the perturbed Hamiltonian

$$H = H_* + \delta\lambda \mathcal{O}, \quad (403)$$

where  $\delta\lambda$  is the perturbation of the coupling constant corresponding to  $\mathcal{O}$  (we can also consider perturbations in more than one coupling constant). At the same time thinking of the  $\lambda_a$ 's as coordinates in the space of all Hamiltonians this corresponds to moving slightly away from the position of the fixed point.

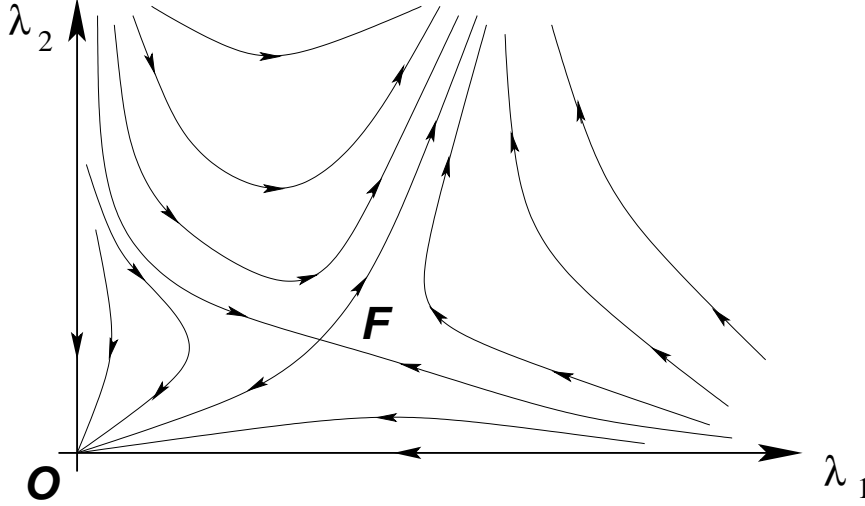
The question to decide now is in which direction the renormalization group flow will take the perturbed system. Working at first order in  $\delta\lambda$  there are three possibilities:

- The renormalization group flow takes the system back to the fixed point. In this case the corresponding interaction  $\mathcal{O}$  is called *irrelevant*.
- $\mathcal{R}$  takes the system away from the fixed point. If this is what happens the interaction is called *relevant*.
- It is possible that the perturbation actually does not take the system away from the fixed point at first order in  $\delta\lambda$ . In this case the interaction is said to be *marginal* and it is necessary to go to higher orders in  $\delta\lambda$  in order to decide whether the system moves to or away the fixed point, or whether we have a family of fixed points.

Therefore we can picture the action of the renormalization group transformation as a flow in the space of coupling constants. In Fig. 17 we have depicted an example of such a flow in the case of a system with two coupling constants  $\lambda_1$  and  $\lambda_2$ . In this example we find two fixed points, one at the origin  $O$  and another at  $F$  for a finite value of the couplings. The arrows indicate the direction in which the renormalization group flow acts. The free theory at  $\lambda_1 = \lambda_2 = 0$  is a stable fix point since any perturbation  $\delta\lambda_1, \delta\lambda_2 > 0$  makes the theory flow back to the free theory at long distances. On the other hand, the fixed point  $F$  is stable with respect to certain type of perturbations (along the line with incoming arrows) whereas for any other perturbations the system flows either to the free theory at the origin or to a theory with infinite values for the couplings.

**Quantum field theory.** Let us see now how these ideas of the renormalization group apply to Field Theory. Let us begin with a quantum field theory defined by the Lagrangian

$$\mathcal{L}[\phi_a] = \mathcal{L}_0[\phi_a] + \sum_i g_i \mathcal{O}_i[\phi_a], \quad (404)$$



**Fig. 17:** Example of a renormalization group flow.

where  $\mathcal{L}_0[\phi_a]$  is the kinetic part of the Lagrangian and  $g_i$  are the coupling constants associated with the operators  $\mathcal{O}_i[\phi_a]$ . In order to make sense of the quantum theory we introduce a cutoff in momenta  $\Lambda$ . In principle we include all operators  $\mathcal{O}_i$  compatible with the symmetries of the theory.

In section 8.2 we saw how in the cases of QED and QCD, the value of the coupling constant changed with the scale from its value at the scale  $\Lambda$ . We can understand now this behavior along the lines of the analysis presented above for the Ising model. If we would like to compute the effective dynamics of the theory at an energy scale  $\mu < \Lambda$  we only have to integrate out all physical models with energies between the cutoff  $\Lambda$  and the scale of interest  $\mu$ . This is analogous to what we did in the Ising model by replacing the original spins by the block spins. In the case of field theory the effective action  $S[\phi_a, \mu]$  at scale  $\mu$  can be written in the language of functional integration as

$$e^{iS[\phi'_a, \mu]} = \int_{\mu < p < \Lambda} \prod_a \mathcal{D}\phi_a e^{iS[\phi_a, \Lambda]}. \quad (405)$$

Here  $S[\phi_a, \Lambda]$  is the action at the cutoff scale

$$S[\phi_a, \Lambda] = \int d^4x \left\{ \mathcal{L}_0[\phi_a] + \sum_i g_i(\Lambda) \mathcal{O}_i[\phi_a] \right\} \quad (406)$$

and the functional integral in Eq. (405) is carried out only over the field modes with momenta in the range  $\mu < p < \Lambda$ . The action resulting from integrating out the physics at the intermediate scales between  $\Lambda$  and  $\mu$  depends not on the original field variable  $\phi_a$  but on some renormalized field  $\phi'_a$ . At the same time the couplings  $g_i(\mu)$  differ from their values at the cutoff scale  $g_i(\Lambda)$ . This is analogous to what we learned in the Ising model: by integrating out short distance physics we ended up with a new Hamiltonian depending on renormalized effective spin variables and with renormalized values for the coupling constants. Therefore the resulting effective action at scale  $\mu$  can be written as

$$S[\phi'_a, \mu] = \int d^4x \left\{ \mathcal{L}_0[\phi'_a] + \sum_i g_i(\mu) \mathcal{O}_i[\phi'_a] \right\}. \quad (407)$$

This Wilsonian interpretation of renormalization sheds light to what in section 8.1 might have looked just a smart way to get rid of the infinities. The running of the coupling constant with the energy scale can be understood now as a way of incorporating into an effective action at scale  $\mu$  the effects of field excitations at higher energies  $E > \mu$ .

As in statistical mechanics there are also quantum field theories that are fixed points of the renormalization group flow, i.e. whose coupling constants do not change with the scale. We have encountered them already in Section 8.2 when studying the properties of the beta function. The most trivial example of such theories are massless free quantum field theories, but there are also examples of four-dimensional interacting quantum field theories which are scale invariant. Again we can ask the question of what happens when a scale invariant theory is perturbed with some operator. In general the perturbed theory is not scale invariant anymore but we may wonder whether the perturbed theory flows at low energies towards or away the theory at the fixed point.

In quantum field theory this can be decided by looking at the canonical dimension  $d[\mathcal{O}]$  of the operator  $\mathcal{O}[\phi_a]$  used to perturb the theory at the fixed point. In four dimensions the three possibilities are defined by:

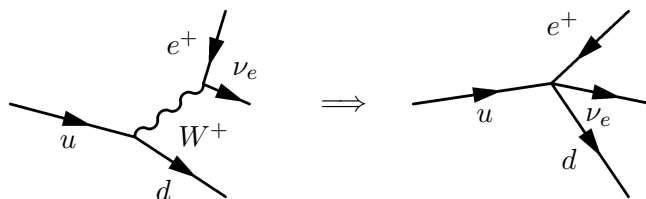
- $d[\mathcal{O}] > 4$ : irrelevant perturbation. The running of the coupling constants takes the theory back to the fixed point.
- $d[\mathcal{O}] < 4$ : relevant perturbation. At low energies the theory flows away from the scale-invariant theory.
- $d[\mathcal{O}] = 4$ : marginal deformation. The direction of the flow cannot be decided only on dimensional grounds.

As an example, let us consider first a massless fermion theory perturbed by a four-fermion interaction term

$$\mathcal{L} = i\bar{\psi}\not{\partial}\psi - \frac{1}{M^2}(\bar{\psi}\psi)^2. \tag{408}$$

This is indeed a perturbation by an irrelevant operator, since in four-dimensions  $[\psi] = \frac{3}{2}$ . Interactions generated by the extra term are suppressed at low energies since typically their effects are weighted by the dimensionless factor  $\frac{E^2}{M^2}$ , where  $E$  is the energy scale of the process. This means that as we try to capture the relevant physics at lower and lower energies the effect of the perturbation is weaker and weaker rendering in the infrared limit  $E \rightarrow 0$  again a free theory. Hence, the irrelevant perturbation in (408) makes the theory flow back to the fixed point.

On the other hand relevant operators dominate the physics at low energies. This is the case, for example, of a mass term. As we lower the energy the mass becomes more important and once the energy goes below the mass of the field its dynamics is completely dominated by the mass term. This is, for example, how Fermi's theory of weak interactions emerges from the standard model at energies below the mass of the  $W^\pm$  boson



At energies below  $M_W = 80.4$  GeV the dynamics of the  $W^+$  boson is dominated by its mass term and therefore becomes nonpropagating, giving rise to the effective four-fermion Fermi theory.

To summarize our discussion so far, we found that while relevant operators dominate the dynamics in the infrared, taking the theory away from the fixed point, irrelevant perturbations become suppressed in the same limit. Finally we consider the effect of marginal operators. As an example we take the interaction term in massless QED,  $\mathcal{O} = \bar{\psi}\gamma^\mu\psi A_\mu$ . Taking into account that in  $d = 4$  the dimension of the electromagnetic potential is  $[A_\mu] = 1$  the operator  $\mathcal{O}$  is a marginal perturbation. In order to decide whether the fixed point theory

$$\mathcal{L}_0 = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\not{D}\psi \tag{409}$$

is restored at low energies or not we need to study the perturbed theory in more detail. This we have done in section 8.1 where we learned that the effective coupling in QED decreases at low energies. Then we conclude that the perturbed theory flows towards the fixed point in the infrared.

As an example of a marginal operator with the opposite behavior we can write the Lagrangian for a  $SU(N_c)$  gauge theory,  $\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^a F^{a\mu\nu}$ , as

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4}(\partial_\mu A_\nu^a - \partial_\nu A_\mu^a)(\partial^\mu A^{a\nu} - \partial^\nu A^{a\mu}) - 4gf^{abc}A_\mu^a A_\nu^b \partial^\mu A^{c\nu} \\ &+ g^2 f^{abc} f^{ade} A_\mu^b A_\nu^c A^{d\mu} A^{e\nu} \equiv \mathcal{L}_0 + \mathcal{O}_g, \end{aligned} \quad (410)$$

i.e. a marginal perturbation of the free theory described by  $\mathcal{L}_0$ , which is obviously a fixed point under renormalization group transformations. Unlike the case of QED we know that the full theory is asymptotically free, so the coupling constant grows at low energies. This implies that the operator  $\mathcal{O}_g$  becomes more and more important in the infrared and therefore the theory flows away the fixed point in this limit.

It is very important to notice here that in the Wilsonian view the cutoff is not necessarily regarded as just some artifact to remove infinities but actually has a physical origin. For example in the case of Fermi's theory of  $\beta$ -decay there is a natural cutoff  $\Lambda = M_W$  at which the theory has to be replaced by the standard model. In the case of the standard model itself the cutoff can be taken at Planck scale  $\Lambda \simeq 10^{19}$  GeV or the Grand Unification scale  $\Lambda \simeq 10^{16}$  GeV, where new degrees of freedom are expected to become relevant. The cutoff serves the purpose of cloaking the range of energies at which new physics has to be taken into account.

Provided that in the Wilsonian approach the quantum theory is always defined with a physical cutoff, there is no fundamental difference between renormalizable and nonrenormalizable theories. Actually, a renormalizable field theory, like the standard model, can generate nonrenormalizable operators at low energies such as the effective four-fermion interaction of Fermi's theory. They are not sources of any trouble if we are interested in the physics at scales much below the cutoff,  $E \ll \Lambda$ , since their contribution to the amplitudes will be suppressed by powers of  $\frac{E}{\Lambda}$ .

## 9 Special topics

### 9.1 Creation of particles by classical fields

**Particle creation by a classical source.** In a free quantum field theory the total number of particles contained in a given state of the field is a conserved quantity. For example, in the case of the quantum scalar field studied in section 3 we have that the number operator commutes with the Hamiltonian

$$\hat{n} \equiv \int \frac{d^3k}{(2\pi)^3} \frac{1}{2\omega_k} \alpha^\dagger(\vec{k})\alpha(\vec{k}), \quad [\hat{H}, \hat{n}] = 0. \quad (411)$$

This means that any states with a well-defined number of particle excitations will preserve this number at all times. The situation, however, changes as soon as interactions are introduced, since in this case particles can be created and/or destroyed as a result of the dynamics.

Another case in which the number of particles might change is if the quantum theory is coupled to a classical source. The archetypical example of such a situation is the Schwinger effect, in which a classical strong electric field produces the creation of electron-positron pairs out of the vacuum. However, before plunging into this more involved situation we can illustrate the relevant physics involved in the creation of particles by classical sources with the help of the simplest example: a free scalar field theory coupled to a classical external source  $J(x)$ . The action for such a theory can be written as

$$S = \int d^4x \left[ \frac{1}{2} \partial_\mu \phi(x) \partial^\mu \phi(x) - \frac{m^2}{2} \phi(x)^2 + J(x) \phi(x) \right], \quad (412)$$

where  $J(x)$  is a real function of the coordinates. Its identification with a classical source is obvious once we calculate the equations of motion

$$(\nabla^2 + m^2) \phi(x) = J(x). \quad (413)$$

Our plan is to quantize this theory but, unlike the case analyzed in section 3, now the presence of the source  $J(x)$  makes the situation a bit more involved. The general solution to the equations of motion can be written in terms of the retarded Green function for the Klein-Gordon equation as

$$\phi(x) = \phi_0(x) + i \int d^4x' G_R(x-x') J(x'), \quad (414)$$

where  $\phi_0(x)$  is a general solution to the homogeneous equation and

$$\begin{aligned} G_R(t, \vec{x}) &= \int \frac{d^4k}{(2\pi)^4} \frac{i}{k^2 - m^2 + i\epsilon \text{sign}(k^0)} e^{-ik \cdot x} \\ &= i \theta(t) \int \frac{d^3k}{(2\pi)^3} \frac{1}{2\omega_k} \left( e^{-i\omega_k t + i\vec{k} \cdot \vec{x}} - e^{i\omega_k t - i\vec{k} \cdot \vec{x}} \right), \end{aligned} \quad (415)$$

with  $\theta(x)$  the Heaviside step function. The integration contour to evaluate the integral over  $p^0$  surrounds the poles at  $p^0 = \pm\omega_k$  from above. Since  $G_R(t, \vec{x}) = 0$  for  $t < 0$ , the function  $\phi_0(x)$  corresponds to the solution of the field equation at  $t \rightarrow -\infty$ , before the interaction with the external source<sup>21</sup>

To make the argument simpler we assume that  $J(x)$  is switched on at  $t = 0$ , and only last for a time  $\tau$ , that is

$$J(t, \vec{x}) = 0 \quad \text{if } t < 0 \text{ or } t > \tau. \quad (416)$$

We are interested in a solution of (413) for times after the external source has been switched off,  $t > \tau$ . In this case the expression (415) can be written in terms of the Fourier modes  $\tilde{J}(\omega, \vec{k})$  of the source as

$$\phi(t, \vec{x}) = \phi_0(x) + i \int \frac{d^3k}{(2\pi)^3} \frac{1}{2\omega_k} \left[ \tilde{J}(\omega_k, \vec{k}) e^{-i\omega_k t + i\vec{k} \cdot \vec{x}} - \tilde{J}(\omega_k, \vec{k})^* e^{i\omega_k t - i\vec{k} \cdot \vec{x}} \right]. \quad (417)$$

On the other hand, the general solution  $\phi_0(x)$  has been already computed in Eq. (77). Combining this result with Eq. (417) we find the following expression for the late time general solution to the Klein-Gordon equation in the presence of the source

$$\begin{aligned} \phi(t, x) &= \int \frac{d^3k}{(2\pi)^3} \frac{1}{\sqrt{2\omega_k}} \left\{ \left[ \alpha(\vec{k}) + \frac{i}{\sqrt{2\omega_k}} \tilde{J}(\omega_k, \vec{k}) \right] e^{-i\omega_k t + i\vec{k} \cdot \vec{x}} \right. \\ &\quad \left. + \left[ \alpha^*(\vec{k}) - \frac{i}{\sqrt{2\omega_k}} \tilde{J}(\omega_k, \vec{k})^* \right] e^{i\omega_k t - i\vec{k} \cdot \vec{x}} \right\}. \end{aligned} \quad (418)$$

We should not forget that this is a solution valid for times  $t > \tau$ , i.e. once the external source has been disconnected. On the other hand, for  $t < 0$  we find from Eqs. (414) and (415) that the general solution is given by Eq. (77).

Now we can proceed to quantize the theory. The conjugate momentum  $\pi(x) = \partial_0 \phi(x)$  can be computed from Eqs. (77) and (418). Imposing the canonical equal time commutation relations (74) we find that  $\alpha(\vec{k})$ ,  $\alpha^\dagger(\vec{k})$  satisfy the creation-annihilation algebra (51). From our previous calculation we find that for  $t > \tau$  the expansion of the operator  $\phi(x)$  in terms of the creation-annihilation operators  $\alpha(\vec{k})$ ,  $\alpha^\dagger(\vec{k})$  can be obtained from the one for  $t < 0$  by the replacement

$$\alpha(\vec{k}) \longrightarrow \beta(\vec{k}) \equiv \alpha(\vec{k}) + \frac{i}{\sqrt{2\omega_k}} \tilde{J}(\omega_k, \vec{k}),$$

<sup>21</sup>We could have taken instead the advanced propagator  $G_A(x)$  in which case  $\phi_0(x)$  would correspond to the solution to the equation at large times, after the interaction with  $J(x)$ .

$$\alpha^\dagger(\vec{k}) \longrightarrow \beta^\dagger(\vec{k}) \equiv \alpha^\dagger(\vec{k}) - \frac{i}{\sqrt{2\omega_k}} \tilde{J}(\omega_k, \vec{k})^*. \quad (419)$$

Actually, since  $\tilde{J}(\omega_k, \vec{k})$  is a c-number, the operators  $\beta(\vec{k}), \beta^\dagger(\vec{k})$  satisfy the same algebra as  $\alpha(\vec{k}), \alpha^\dagger(\vec{k})$  and therefore can be interpreted as well as a set of creation-annihilation operators. This means that we can define two vacuum states,  $|0_-\rangle, |0_+\rangle$  associated with both sets of operators

$$\left. \begin{array}{l} \alpha(\vec{k})|0_-\rangle = 0 \\ \beta(\vec{k})|0_+\rangle = 0 \end{array} \right\} \quad \forall \vec{k}. \quad (420)$$

For an observer at  $t < 0$ ,  $\alpha(\vec{k})$  and  $\alpha^\dagger(\vec{k})$  are the natural set of creation-annihilation operators in terms of which to expand the field operator  $\phi(x)$ . After the usual zero-point energy subtraction the Hamiltonian is given by

$$\hat{H}^{(-)} = \frac{1}{2} \int \frac{d^3k}{(2\pi)^3} \alpha^\dagger(\vec{k}) \alpha(\vec{k}) \quad (421)$$

and the ground state of the spectrum for this observer is the vacuum  $|0_-\rangle$ . At the same time, a second observer at  $t > \tau$  will also see a free scalar quantum field (the source has been switched off at  $t = \tau$ ) and consequently will expand  $\phi$  in terms of the second set of creation-annihilation operators  $\beta(\vec{k}), \beta^\dagger(\vec{k})$ . In terms of this operators the Hamiltonian is written as

$$\hat{H}^{(+)} = \frac{1}{2} \int \frac{d^3k}{(2\pi)^3} \beta^\dagger(\vec{k}) \beta(\vec{k}). \quad (422)$$

Then for this late-time observer the ground state of the Hamiltonian is the second vacuum state  $|0_+\rangle$ .

In our analysis we have been working in the Heisenberg picture, where states are time-independent and the time dependence comes in the operators. Therefore the states of the theory are globally defined. Suppose now that the system is in the “in” ground state  $|0_-\rangle$ . An observer at  $t < 0$  will find that there are no particles

$$\hat{n}^{(-)}|0_-\rangle = 0. \quad (423)$$

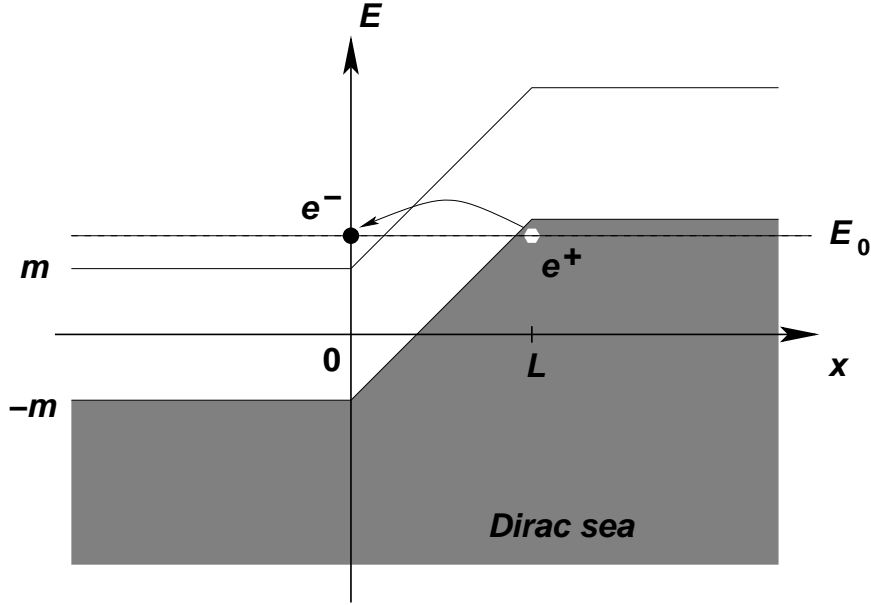
However the late-time observer will find that the state  $|0_-\rangle$  contains an average number of particles given by

$$\langle 0_- | \hat{n}^{(+)} | 0_- \rangle = \int \frac{d^3k}{(2\pi)^3} \frac{1}{2\omega_k} \left| \tilde{J}(\omega_k, \vec{k}) \right|^2. \quad (424)$$

Moreover,  $|0_-\rangle$  is no longer the ground state for the “out” observer. On the contrary, this state have a vacuum expectation value for  $\hat{H}^{(+)}$

$$\langle 0_- | \hat{H}^{(+)} | 0_- \rangle = \frac{1}{2} \int \frac{d^3k}{(2\pi)^3} \left| \tilde{J}(\omega_k, \vec{k}) \right|^2. \quad (425)$$

The key to understand what is going on here lies in the fact that the external source breaks the invariance of the theory under space-time translations. In the particular case we have studied here where  $J(x)$  has support over a finite time interval  $0 < t < \tau$ , this implies that the vacuum is not invariant under time translations, so observers at different times will make different choices of vacuum that will not necessarily agree with each other. This is clear in our example. An observer in  $t < \tau$  will choose the vacuum to be the lowest energy state of her Hamiltonian,  $|0_-\rangle$ . On the other hand, the second observer at late times  $t > \tau$  will naturally choose  $|0_+\rangle$  as the vacuum. However, for this second observer, the



**Fig. 18:** Pair creation by a electric field in the Dirac sea picture.

state  $|0_-\rangle$  is not the vacuum of his Hamiltonian, but actually an excited state that is a superposition of states with well-defined number of particles. In this sense it can be said that the external source has the effect of creating particles out of the “in” vacuum. Besides, this breaking of time translation invariance produces a violation in the energy conservation as we see from Eq. (425). Particles are actually created from the energy pumped into the system by the external source.

**The Schwinger effect.** A classical example of creation of particles by a external field was pointed out by Schwinger [44] and consists of the creation of electron-positron pairs by a strong electric field. In order to illustrate this effect we are going to follow a heuristic argument based on the Dirac sea picture and the WKB approximation.

In the absence of an electric field the vacuum state of a spin- $\frac{1}{2}$  field is constructed by filling all the negative energy states as depicted in Fig. 2. Let us now connect a constant electric field  $\vec{\mathcal{E}} = \mathcal{E}\vec{u}_x$  in the range  $0 < x < L$  created by a electrostatic potential

$$V(\vec{r}) = \begin{cases} 0 & x < 0 \\ -\mathcal{E}x & 0 < x < L \\ -\mathcal{E}L & x > L \end{cases} \quad (426)$$

After the field has been switched on, the Dirac sea looks like in Fig. 18. In particular we find that if  $e\mathcal{E}L > 2m$  there are negative energy states at  $x > L$  with the same energy as the positive energy states in the region  $x < 0$ . Therefore it is possible for an electron filling a negative energy state with energy close to  $-2m$  to tunnel through the forbidden region into a positive energy state. The interpretation of such a process is the production of an electron-positron pair out of the electric field.

We can compute the rate at which such pairs are produced by using the WKB approximation. Focusing for simplicity on an electron on top of the Fermi surface near  $x = L$  with energy  $E_0$ , the transmission coefficient in this approximation is given by<sup>22</sup>

$$T_{\text{WKB}} = \exp \left[ -2 \int_{\frac{1}{e\mathcal{E}}(E_0 - \sqrt{m^2 + \vec{p}_T^2})}^{\frac{1}{e\mathcal{E}}(E_0 + \sqrt{m^2 + \vec{p}_T^2})} dx \sqrt{m^2 - [E_0 - e\mathcal{E}(x - x_0)]^2 + \vec{p}_T^2} \right]$$

<sup>22</sup>Notice that the electron satisfy the relativistic dispersion relation  $E = \sqrt{\vec{p}^2 + m^2} + V$  and therefore  $-\vec{p}_x^2 = m^2 - (E - V)^2 + \vec{p}_T^2$ . The integration limits are set by those values of  $x$  at which  $p_x = 0$ .

$$= \exp \left[ -\frac{\pi}{e\mathcal{E}} (\vec{p}_T^2 + m^2) \right], \quad (427)$$

where  $p_T^2 \equiv p_y^2 + p_z^2$ . This gives the transition probability per unit time and per unit cross section  $dydz$  for an electron in the Dirac sea with transverse momentum  $\vec{p}_T$  and energy  $E_0$ . To get the total probability per unit time and per unit volume we have to integrate over all possible values of  $\vec{p}_T$  and  $E_0$ . Actually, in the case of the energy, because of the relation between  $E_0$  and the coordinate  $x$  at which the particle penetrates into the barrier we can write  $\frac{dE_0}{2\pi} = \frac{e\mathcal{E}}{2\pi} dx$  and the total probability per unit time and per unit volume for the creation of a pair is given by

$$W = 2 \left( \frac{e\mathcal{E}}{2\pi} \right) \int \frac{d^2 p_T}{(2\pi)^2} e^{-\frac{\pi}{e\mathcal{E}} (\vec{p}_T^2 + m^2)} = \frac{e^2 \mathcal{E}^2}{4\pi^3} e^{-\frac{\pi m^2}{e\mathcal{E}}}, \quad (428)$$

where the factor of 2 accounts for the two polarizations of the electron.

Then production of electron-positron pairs is exponentially suppressed and it is only sizeable for strong electric fields. To estimate its order of magnitude it is useful to restore the powers of  $c$  and  $\hbar$  in (428)

$$W = \frac{e^2 \mathcal{E}^2}{4\pi^3 c \hbar^2} e^{-\frac{\pi m^2 c^3}{\hbar e \mathcal{E}}} \quad (429)$$

The exponential suppression of the pair production disappears when the electric field reaches the critical value  $\mathcal{E}_{\text{crit}}$  at which the exponent is of order one

$$\mathcal{E}_{\text{crit}} = \frac{m^2 c^3}{\hbar e} \simeq 1.3 \times 10^{16} \text{ V cm}^{-1}. \quad (430)$$

This is indeed a very strong field which is extremely difficult to produce. A similar effect, however, takes place also in a time-varying electric field [45] and there is the hope that pair production could be observed in the presence of the alternating electric field produced by a laser.

The heuristic derivation that we followed here can be made more precise in QED. There the decay of the vacuum into electron-positron pairs can be computed from the imaginary part of the effective action  $\Gamma[A_\mu]$  in the presence of a classical gauge potential  $A_\mu$

$$\begin{aligned} i\Gamma[A_\mu] &\equiv \text{Diagram 1} + \text{Diagram 2} + \text{Diagram 3} + \dots \\ &= \log \det \left[ 1 - ieA \frac{1}{i\cancel{\partial} - m} \right]. \end{aligned} \quad (431)$$

This determinant can be computed using the standard heat kernel techniques. The probability of pair production is proportional to the imaginary part of  $i\Gamma[A_\mu]$  and gives

$$W = \frac{e^2 \mathcal{E}^2}{4\pi^3} \sum_{n=1}^{\infty} \frac{1}{n^2} e^{-n \frac{\pi m^2}{e\mathcal{E}}}. \quad (432)$$

Our simple argument based on tunneling in the Dirac sea gave only the leading term of Schwinger's result (432). The remaining terms can be also captured in the WKB approximation by taking into account the probability of production of several pairs, i.e. the tunneling of more than one electron through the barrier.

Here we have illustrated the creation of particles by semiclassical sources in quantum field theory using simple examples. Nevertheless, what we learned has important applications to the study of quantum fields in curved backgrounds. In quantum field theory in Minkowski space-time the vacuum state



is invariant under the Poincaré group and this, together with the covariance of the theory under Lorentz transformations, implies that all inertial observers agree on the number of particles contained in a quantum state. The breaking of such invariance, as happened in the case of coupling to a time-varying source analyzed above, implies that it is not possible anymore to define a state which would be recognized as the vacuum by all observers.

This is precisely the situation when fields are quantized on curved backgrounds. In particular, if the background is time-dependent (as it happens in a cosmological setup or for a collapsing star) different observers will identify different vacuum states. As a consequence what one observer call the vacuum will be full of particles for a different observer. This is precisely what is behind the phenomenon of Hawking radiation [46]. The emission of particles by a physical black hole formed from gravitational collapse of a star is the consequence of the fact that the vacuum state in the asymptotic past contain particles for an observer in the asymptotic future. As a consequence, a detector located far away from the black hole detects a stream of thermal radiation with temperature

$$T_{\text{Hawking}} = \frac{\hbar c^3}{8\pi G_N k M} \quad (433)$$

where  $M$  is the mass of the black hole,  $G_N$  is Newton's constant and  $k$  is Boltzmann's constant. There are several ways in which this results can be obtained. A more heuristic way is perhaps to think of this particle creation as resulting from quantum tunneling of particles across the potential barrier posed by gravity [47].

## 9.2 Supersymmetry

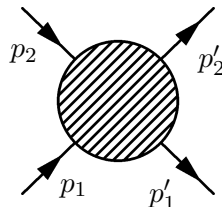
One of the things that we have learned in our journey around the landscape of quantum field theory is that our knowledge of the fundamental interactions in Nature is based on the idea of symmetry, and in particular gauge symmetry. The Lagrangian of the standard model can be written just including all possible renormalizable terms (i.e. with canonical dimension smaller or equal to 4) compatible with the gauge symmetry  $SU(3) \times SU(2) \times U(1)$  and Poincaré invariance. All attempts to go beyond start with the question of how to extend the symmetries of the standard model.

As explained in Section 5.1, in a quantum field theoretical description of the interaction of elementary particles the basic observable quantity to compute is the scattering or  $S$ -matrix giving the probability amplitude for the scattering of a number of incoming particles with a certain momentum into some final products

$$\mathcal{A}(\text{in} \longrightarrow \text{out}) = \langle \vec{p}'_1, \dots; \text{out} | \vec{p}_1, \dots; \text{in} \rangle. \quad (434)$$

An explicit symmetry of the theory has to be necessarily a symmetry of the  $S$ -matrix. Hence it is fair to ask what is the largest symmetry of the  $S$ -matrix.

Let us ask this question in the simple case of the scattering of two particles with four-momenta  $p_1$  and  $p_2$  in the  $t$ -channel



We will make the usual assumptions regarding positivity of the energy and analyticity. Invariance of the theory under the Poincaré group implies that the amplitude can only depend on the scattering angle  $\vartheta$  through

$$t = (p'_1 - p_1)^2 = 2(m_1^2 - p_1 \cdot p'_1) = 2(m_1^2 - E_1 E'_1 + |\vec{p}_1| |\vec{p}'_1| \cos \vartheta). \quad (435)$$

If there would be any extra bosonic symmetry of the theory it would restrict the scattering angle to a set of discrete values. In this case the  $S$ -matrix cannot be analytic since it would vanish everywhere except for the discrete values selected by the extra symmetry.

Actually, the only way to extend the symmetry of the theory without renouncing to the analyticity of the scattering amplitudes is to introduce “fermionic” symmetries, i.e. symmetries whose generators are anticommuting objects [48]. This means that in addition to the generators of the Poincaré group<sup>23</sup>  $P^\mu$ ,  $M^{\mu\nu}$  and the ones for the internal gauge symmetries  $G$ , we can introduce a number of fermionic generators  $Q_a^I$ ,  $\bar{Q}_{\dot{a}I}$  ( $I = 1, \dots, \mathcal{N}$ ), where  $\bar{Q}_{\dot{a}I} = (Q_a^I)^\dagger$ . The most general algebra that these generators satisfy is the  $\mathcal{N}$ -extended supersymmetry algebra [49]

$$\begin{aligned} \{Q_a^I, \bar{Q}_{\dot{b}J}\} &= 2\sigma_{ab}^\mu P_\mu \delta^I_J, \\ \{Q_a^I, Q_b^J\} &= 2\varepsilon_{ab} \mathcal{Z}^{IJ}, \end{aligned} \quad (436)$$

$$\{\bar{Q}_{\dot{a}I}, \bar{Q}_{\dot{b}J}\} = 2\varepsilon_{\dot{a}\dot{b}} \bar{\mathcal{Z}}^{IJ}, \quad (437)$$

where  $\mathcal{Z}^{IJ} \in \mathbb{C}$  commute with any other generator and satisfies  $\mathcal{Z}^{IJ} = -\mathcal{Z}^{JI}$ . Besides we have the commutators that determine the Poincaré transformations of the fermionic generators  $Q_a^I$ ,  $\bar{Q}_{\dot{a}I}$

$$\begin{aligned} [Q_a^I, P^\mu] &= [\bar{Q}_{\dot{a}I}, P^\mu] = 0, \\ [Q_a^I, M^{\mu\nu}] &= \frac{1}{2}(\sigma^{\mu\nu})_a{}^b Q_b^I, \\ [\bar{Q}_{\dot{a}I}, M^{\mu\nu}] &= -\frac{1}{2}(\bar{\sigma}^{\mu\nu})_{\dot{a}}{}^{\dot{b}} \bar{Q}_{\dot{b}I}, \end{aligned} \quad (438)$$

where  $\sigma^{0i} = -i\sigma^i$ ,  $\sigma^{ij} = \varepsilon^{ijk}\sigma^k$  and  $\bar{\sigma}^{\mu\nu} = (\sigma^{\mu\nu})^\dagger$ . These identities simply mean that  $Q_a^I$ ,  $\bar{Q}_{\dot{a}I}$  transform respectively in the  $(\frac{1}{2}, \mathbf{0})$  and  $(\mathbf{0}, \frac{1}{2})$  representations of the Lorentz group.

We know that the presence of a global symmetry in a theory implies that the spectrum can be classified in multiplets with respect to that symmetry. In the case of supersymmetry start with the case  $\mathcal{N} = 1$  in which there is a single pair of supercharges  $Q_a$ ,  $\bar{Q}_{\dot{a}}$  satisfying the algebra

$$\{Q_a, \bar{Q}_{\dot{b}}\} = 2\sigma_{ab}^\mu P_\mu, \quad \{Q_a, Q_b\} = \{\bar{Q}_{\dot{a}}, \bar{Q}_{\dot{b}}\} = 0. \quad (439)$$

Notice that in the  $\mathcal{N} = 1$  case there is no possibility of having central charges.

We study now the representations of the supersymmetry algebra (439), starting with the massless case. Given a state  $|k\rangle$  satisfying  $k^2 = 0$ , we can always find a reference frame where the four-vector  $k^\mu$  takes the form  $k^\mu = (E, 0, 0, E)$ . Since the theory is Lorentz covariant we can obtain the representation of the supersymmetry algebra in this frame where the expressions are simpler. In particular, the right-hand side of the first anticommutator in Eq. (439) is given by

$$2\sigma_{ab}^\mu P_\mu = 2(P^0 - \sigma^3 P^3) = \begin{pmatrix} 0 & 0 \\ 0 & 4E \end{pmatrix}. \quad (440)$$

Therefore the algebra of supercharges in the massless case reduces to

$$\begin{aligned} \{Q_1, Q_1^\dagger\} &= \{Q_1, Q_2^\dagger\} = 0, \\ \{Q_2, Q_2^\dagger\} &= 4E. \end{aligned} \quad (441)$$

The commutator  $\{Q_1, Q_1^\dagger\} = 0$  implies that the action of  $Q_1$  on any state gives a zero-norm state of the Hilbert space  $\|Q_1|\Psi\rangle\| = 0$ . If we want the theory to preserve unitarity we must eliminate these null

<sup>23</sup>The generators  $M^{\mu\nu}$  are related with the ones for boost and rotations introduced in section 4.1 by  $J^i \equiv M^{0i}$ ,  $M^i = \frac{1}{2}\varepsilon^{ijk}M^{jk}$ . In this section we also use the “dotted spinor” notation, in which spinors in the  $(\frac{1}{2}, \mathbf{0})$  and  $(\mathbf{0}, \frac{1}{2})$  representations of the Lorentz group are indicated respectively by undotted  $(a, b, \dots)$  and dotted  $(\dot{a}, \dot{b}, \dots)$  indices.

states from the spectrum. This is equivalent to setting  $Q_1 \equiv 0$ . On the other hand, in terms of the second generator  $Q_2$  we can define the operators

$$a = \frac{1}{2\sqrt{E}}Q_2, \quad a^\dagger = \frac{1}{2\sqrt{E}}Q_2^\dagger, \quad (442)$$

which satisfy the algebra of a pair of fermionic creation-annihilation operators,  $\{a, a^\dagger\} = 1$ ,  $a^2 = (a^\dagger)^2 = 0$ . Starting with a vacuum state  $a|\lambda\rangle = 0$  with helicity  $\lambda$  we can build the massless multiplet

$$|\lambda\rangle, \quad |\lambda + \frac{1}{2}\rangle \equiv a^\dagger|\lambda\rangle. \quad (443)$$

Here we consider two important cases:

- Scalar multiplet: we take the vacuum state to have zero helicity  $|0^+\rangle$  so the multiplet consists of a scalar and a helicity- $\frac{1}{2}$  state

$$|0^+\rangle, \quad |\frac{1}{2}\rangle \equiv a^\dagger|0^+\rangle. \quad (444)$$

However, this multiplet is not invariant under the CPT transformation which reverses the sign of the helicity of the states. In order to have a CPT-invariant theory we have to add to this multiplet its CPT-conjugate which can be obtained from a vacuum state with helicity  $\lambda = -\frac{1}{2}$

$$|0^-\rangle, \quad |-\frac{1}{2}\rangle. \quad (445)$$

Putting them together we can combine the two zero helicity states with the two fermionic ones into the degrees of freedom of a complex scalar field and a Weyl (or Majorana) spinor.

- Vector multiplet: now we take the vacuum state to have helicity  $\lambda = \frac{1}{2}$ , so the multiplet contains also a massless state with helicity  $\lambda = 1$

$$|\frac{1}{2}\rangle, \quad |1\rangle \equiv a^\dagger|\frac{1}{2}\rangle. \quad (446)$$

As with the scalar multiplet we add the CPT conjugated obtained from a vacuum state with helicity  $\lambda = -1$

$$|-\frac{1}{2}\rangle, \quad |-1\rangle, \quad (447)$$

which together with (446) give the propagating states of a gauge field and a spin- $\frac{1}{2}$  gaugino.

In both cases we see the trademark of supersymmetric theories: the number of bosonic and fermionic states within a multiplet are the same.

In the case of extended supersymmetry we have to repeat the previous analysis for each supersymmetry charge. At the end, we have  $\mathcal{N}$  sets of fermionic creation-annihilation operators  $\{a^I, a_I^\dagger\} = \delta^I_J$ ,  $(a_I)^2 = (a_I^\dagger)^2 = 0$ . Let us work out the case of  $\mathcal{N} = 8$  supersymmetry. Since for several reasons we do not want to have states with helicity larger than 2, we start with a vacuum state  $|-2\rangle$  of helicity  $\lambda = -2$ . The rest of the states of the supermultiplet are obtained by applying the eight different creation operators  $a_I^\dagger$  to the vacuum:

$$\begin{aligned} \lambda = 2 : & \quad a_1^\dagger \dots a_8^\dagger |-2\rangle & \binom{8}{8} = 1 \text{ state,} \\ \lambda = \frac{3}{2} : & \quad a_{I_1}^\dagger \dots a_{I_7}^\dagger |-2\rangle & \binom{8}{7} = 8 \text{ states,} \\ \lambda = 1 : & \quad a_{I_1}^\dagger \dots a_{I_6}^\dagger |-2\rangle & \binom{8}{6} = 28 \text{ states,} \end{aligned}$$

$$\begin{aligned}
 \lambda = \frac{1}{2} : & \quad a_{I_1}^\dagger \dots a_{I_5}^\dagger | - 2 \rangle & \quad \binom{8}{5} = 56 \text{ states,} \\
 \lambda = 0 : & \quad a_{I_1}^\dagger \dots a_{I_4}^\dagger | - 2 \rangle & \quad \binom{8}{4} = 70 \text{ states,} \\
 \lambda = -\frac{1}{2} : & \quad a_{I_1}^\dagger a_{I_2}^\dagger a_{I_3}^\dagger | - 2 \rangle & \quad \binom{8}{3} = 56 \text{ states,} \\
 \lambda = -1 : & \quad a_{I_1}^\dagger a_{I_2}^\dagger | - 2 \rangle & \quad \binom{8}{2} = 28 \text{ states,} \\
 \lambda = -\frac{3}{2} : & \quad a_{I_1}^\dagger | - 2 \rangle & \quad \binom{8}{1} = 8 \text{ states,} \\
 \lambda = -2 : & \quad | - 2 \rangle & \quad 1 \text{ state.}
 \end{aligned} \tag{448}$$

Putting together the states with opposite helicity we find that the theory contains:

- 1 spin-2 field  $g_{\mu\nu}$  (a graviton),
- 8 spin- $\frac{3}{2}$  gravitino fields  $\psi_\mu^I$ ,
- 28 gauge fields  $A_\mu^{[IJ]}$ ,
- 56 spin- $\frac{1}{2}$  fermions  $\psi^{[IJK]}$ ,
- 70 scalars  $\phi^{[IJKL]}$ ,

where by  $[IJ\dots]$  we have denoted that the indices are antisymmetrized. We see that, unlike the massless multiplets of  $\mathcal{N} = 1$  supersymmetry studied above, this multiplet is CPT invariant by itself. As in the case of the massless  $\mathcal{N} = 1$  multiplet, here we also find as many bosonic as fermionic states:

$$\begin{aligned}
 \text{bosons:} & \quad 1 + 28 + 70 + 28 + 1 = 128 \quad \text{states,} \\
 \text{fermions:} & \quad 8 + 56 + 56 + 8 = 128 \quad \text{states.}
 \end{aligned}$$

Now we study briefly the case of massive representations  $|k\rangle$ ,  $k^2 = M^2$ . Things become simpler if we work in the rest frame where  $P^0 = M$  and the spatial components of the momentum vanish. Then, the supersymmetry algebra becomes:

$$\{Q_a^I, \bar{Q}_{bJ}\} = 2M\delta_{ab}\delta^I_J. \tag{449}$$

We proceed now in a similar way to the massless case by defining the operators

$$a_a^I \equiv \frac{1}{\sqrt{2M}} Q_a^I, \quad a_{\dot{a}I}^\dagger \equiv \frac{1}{\sqrt{2M}} \bar{Q}_{\dot{a}I}. \tag{450}$$

The multiplets are found by choosing a vacuum state with a definite spin. For example, for  $\mathcal{N} = 1$  and taking a spin-0 vacuum  $|0\rangle$  we find three states in the multiplet transforming irreducibly with respect to the Lorentz group:

$$|0\rangle, \quad a_{\dot{a}}^\dagger |0\rangle, \quad \varepsilon^{\dot{a}\dot{b}} a_{\dot{a}}^\dagger a_{\dot{b}}^\dagger |0\rangle, \tag{451}$$

which, once transformed back from the rest frame, correspond to the physical states of two spin-0 bosons and one spin- $\frac{1}{2}$  fermion. For  $\mathcal{N}$ -extended supersymmetry the corresponding multiplets can be worked out in a similar way.

The equality between bosonic and fermionic degrees of freedom is at the root of many of the interesting properties of supersymmetric theories. For example, in section 4 we computed the divergent vacuum energy contributions for each real bosonic or fermionic propagating degree of freedom is<sup>24</sup>

$$E_{\text{vac}} = \pm \frac{1}{2} \delta(\vec{0}) \int d^3p \omega_p, \tag{452}$$

<sup>24</sup>For a boson, this can be read off Eq. (80). In the case of fermions, the result of Eq. (134) gives the vacuum energy contribution of the four real propagating degrees of freedom of a Dirac spinor.

where the sign  $\pm$  corresponds respectively to bosons and fermions. Hence, for a supersymmetric theory the vacuum energy contribution exactly cancels between bosons and fermions. This boson-fermion degeneracy is also responsible for supersymmetric quantum field theories being less divergent than non-supersymmetric ones.

### Appendix: A crash course in Group Theory

In this Appendix we summarize some basic facts about Group Theory. Given a group  $G$  a representation of  $G$  is a correspondence between the elements of  $G$  and the set of linear operators acting on a vector space  $V$ , such that for each element of the group  $g \in G$  there is a linear operator  $D(g)$

$$D(g) : V \longrightarrow V \quad (453)$$

satisfying the group operations

$$D(g_1)D(g_2) = D(g_1g_2), \quad D(g_1^{-1}) = D(g_1)^{-1}, \quad g_1, g_2 \in \mathcal{G}. \quad (454)$$

The representation  $D(g)$  is irreducible if and only if the only operators  $A : V \rightarrow V$  commuting with all the elements of the representation  $D(g)$  are the ones proportional to the identity

$$[D(g), A] = 0, \quad \forall g \quad \iff \quad A = \lambda \mathbf{1}, \quad \lambda \in \mathbb{C} \quad (455)$$

More intuitively, we can say that a representation is irreducible if there is no proper subspace  $U \subset V$  (i.e.  $U \neq V$  and  $U \neq \emptyset$ ) such that  $D(g)U \subset U$  for every element  $g \in G$ .

Here we are specially interested in Lie groups whose elements are labelled by a number of continuous parameters. In mathematical terms this means that a Lie group is a manifold  $\mathcal{M}$  together with an operation  $\mathcal{M} \times \mathcal{M} \rightarrow \mathcal{M}$  that we will call multiplication that satisfies the associativity property  $g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3$  together with the existence of unity  $g\mathbf{1} = \mathbf{1}g = g$ , for every  $g \in \mathcal{M}$  and inverse  $gg^{-1} = g^{-1}g = \mathbf{1}$ .

The simplest example of a Lie group is  $\text{SO}(2)$ , the group of rotations in the plane. Each element  $R(\theta)$  is labelled by the rotation angle  $\theta$ , with the multiplication acting as  $R(\theta_1)R(\theta_2) = R(\theta_1 + \theta_2)$ . Because the angle  $\theta$  is defined only modulo  $2\pi$ , the manifold of  $\text{SO}(2)$  is a circumference  $S^1$ .

One of the interesting properties of Lie groups is that in a neighborhood of the identity element they can be expressed in terms of a set of generators  $T^a$  ( $a = 1, \dots, \dim G$ ) as

$$D(g) = \exp(-i\alpha_a T^a) \equiv \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \alpha_{a_1} \dots \alpha_{a_n} T^{a_1} \dots T^{a_n}, \quad (456)$$

where  $\alpha_a \in \mathbb{C}$  are a set of coordinates of  $\mathcal{M}$  in a neighborhood of  $\mathbf{1}$ . Because of the general Baker-Campbell-Hausdorff formula, the multiplication of two group elements is encoded in the value of the commutator of two generators, that in general has the form

$$[T^a, T^b] = if^{abc}T^c, \quad (457)$$

where  $f^{abc} \in \mathbb{C}$  are called the structure constants. The set of generators with the commutator operation form the Lie algebra associated with the Lie group. Hence, given a representation of the Lie algebra of generators we can construct a representation of the group by exponentiation (at least locally near the identity).

We illustrate these concept with some particular examples. For  $\text{SU}(2)$  each group element is labelled by three real number  $\alpha_i$ ,  $i = 1, 2, 3$ . We have two basic representations: one is the fundamental representation (or spin  $\frac{1}{2}$ ) defined by

$$D_{\frac{1}{2}}(\alpha_i) = e^{-\frac{i}{2}\alpha_i\sigma^i}, \quad (458)$$

with  $\sigma^i$  the Pauli matrices. The second one is the adjoint (or spin 1) representation which can be written as

$$D_1(\alpha_i) = e^{-i\alpha_i J^i}, \quad (459)$$

where

$$J^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad J^2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad J^3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \quad (460)$$

Actually,  $J^i$  ( $i = 1, 2, 3$ ) generate rotations around the  $x$ ,  $y$  and  $z$  axis respectively. Representations of spin  $j \in \mathbb{N} + \frac{1}{2}$  can be also constructed with dimension

$$\dim D_j(g) = 2j + 1. \quad (461)$$

As a second example we consider SU(3). This group has two basic three-dimensional representations denoted by  $\mathbf{3}$  and  $\bar{\mathbf{3}}$  which in QCD are associated with the transformation of quarks and antiquarks under the color gauge symmetry SU(3). The elements of these representations can be written as

$$D_{\mathbf{3}}(\alpha^a) = e^{\frac{i}{2}\alpha^a \lambda_a}, \quad D_{\bar{\mathbf{3}}}(\alpha^a) = e^{-\frac{i}{2}\alpha^a \lambda_a^T} \quad (a = 1, \dots, 8), \quad (462)$$

where  $\lambda_a$  are the eight hermitian Gell-Mann matrices

$$\begin{aligned} \lambda_1 &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, & \lambda_2 &= \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, & \lambda_3 &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\ \lambda_4 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, & \lambda_5 &= \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, & \lambda_6 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \\ \lambda_7 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, & \lambda_8 &= \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & -\frac{2}{\sqrt{3}} \end{pmatrix}. \end{aligned} \quad (463)$$

Hence the generators of the representations  $\mathbf{3}$  and  $\bar{\mathbf{3}}$  are given by

$$T^a(\mathbf{3}) = \frac{1}{2}\lambda_a, \quad T^a(\bar{\mathbf{3}}) = -\frac{1}{2}\lambda_a^T. \quad (464)$$

Irreducible representations can be classified in three groups: real, complex and pseudoreal.

- Real representations: a representation is said to be real if there is a *symmetric matrix*  $S$  which acts as intertwiner between the generators and their complex conjugates

$$\bar{T}^a = -S T^a S^{-1}, \quad S^T = S. \quad (465)$$

This is for example the case of the adjoint representation of SU(2) generated by the matrices (460)

- Pseudoreal representations: are the ones for which an *antisymmetric matrix*  $S$  exists with the property

$$\bar{T}^a = -S T^a S^{-1}, \quad S^T = -S. \quad (466)$$

As an example we can mention the spin- $\frac{1}{2}$  representation of SU(2) generated by  $\frac{1}{2}\sigma^i$ .

- Complex representations: finally, a representation is complex if the generators and their complex conjugate are not related by a similarity transformation. This is for instance the case of the two three-dimensional representations  $\mathbf{3}$  and  $\bar{\mathbf{3}}$  of  $SU(3)$ .

There are a number of invariants that can be constructed associated with an irreducible representation  $R$  of a Lie group  $G$  and that can be used to label such a representation. If  $T_R^a$  are the generators in a certain representation  $R$  of the Lie algebra, it is easy to see that the matrix  $\sum_{a=1}^{\dim G} T_R^a T_R^a$  commutes with every generator  $T_R^a$ . Therefore, because of Schur's lemma, it has to be proportional to the identity<sup>25</sup>. This defines the Casimir invariant  $C_2(R)$  as

$$\sum_{a=1}^{\dim G} T_R^a T_R^a = C_2(R) \mathbf{1}. \quad (467)$$

A second invariant  $T_2(R)$  associated with a representation  $R$  can also be defined by the identity

$$\text{Tr } T_R^a T_R^b = T_2(R) \delta^{ab}. \quad (468)$$

Actually, taking the trace in Eq. (467) and combining the result with (468) we find that both invariants are related by the identity

$$C_2(R) \dim R = T_2(R) \dim G, \quad (469)$$

with  $\dim R$  the dimension of the representation  $R$ .

These two invariants appear frequently in quantum field theory calculations with nonabelian gauge fields. For example  $T_2(R)$  comes about as the coefficient of the one-loop calculation of the beta-function for a Yang-Mills theory with gauge group  $G$ . In the case of  $SU(N)$ , for the fundamental representation, we find the values

$$C_2(\text{fund}) = \frac{N^2 - 1}{2N}, \quad T_2(\text{fund}) = \frac{1}{2}, \quad (470)$$

whereas for the adjoint representation the results are

$$C_2(\text{adj}) = N, \quad T_2(\text{adj}) = N. \quad (471)$$

A third invariant  $A(R)$  is specially important in the calculation of anomalies. As discussed in section (7), the chiral anomaly in gauge theories is proportional to the group-theoretical factor  $\text{Tr} [T_R^a \{T_R^b, T_R^c\}]$ . This leads us to define  $A(R)$  as

$$\text{Tr} [T_R^a \{T_R^b, T_R^c\}] = A(R) d^{abc}, \quad (472)$$

where  $d^{abc}$  is symmetric in its three indices and does not depend on the representation. Therefore, the cancellation of anomalies in a gauge theory with fermions transformed in the representation  $R$  of the gauge group is guaranteed if the corresponding invariant  $A(R)$  vanishes.

It is not difficult to prove that  $A(R) = 0$  if the representation  $R$  is either real or pseudoreal. Indeed, if this is the case, then there is a matrix  $S$  (symmetric or antisymmetric) that intertwines the generators  $T_R^a$  and their complex conjugates  $\bar{T}_R^a = -S T_R^a S^{-1}$ . Then, using the hermiticity of the generators we can write

$$\text{Tr} [T_R^a \{T_R^b, T_R^c\}] = \text{Tr} [T_R^a \{T_R^b, T_R^c\}]^T = \text{Tr} [\bar{T}_R^a \{\bar{T}_R^b, \bar{T}_R^c\}]. \quad (473)$$

<sup>25</sup>Schur's lemma states that if there is a matrix  $A$  that commutes with all elements of an irreducible representation of a Lie algebra, then  $A = \lambda \mathbf{1}$ , for some  $\lambda \in \mathbb{C}$ .

Now, using (465) or (466) we have

$$\mathrm{Tr} \left[ \bar{T}_R^a \{ \bar{T}_R^b, \bar{T}_R^c \} \right] = -\mathrm{Tr} \left[ S T_R^a S^{-1} \{ S T_R^b S^{-1}, S T_R^c S^{-1} \} \right] = -\mathrm{Tr} \left[ T_R^a \{ T_R^b, T_R^c \} \right], \quad (474)$$

which proves that  $\mathrm{Tr} \left[ T_R^a \{ T_R^b, T_R^c \} \right]$  and therefore  $A(R) = 0$  whenever the representation is real or pseudoreal. Since the gauge anomaly in four dimensions is proportional to  $A(R)$  this means that anomalies appear only when the fermions transform in a complex representation of the gauge group.

## References

- [1] L. Álvarez-Gaumé and M. A. Vázquez-Mozo, *An Invitation to Quantum Field Theory*, Springer 2011.
- [2] J. D. Bjorken and S. D. Drell, *Relativistic Quantum Fields*, McGraw-Hill 1965.
- [3] C. Itzykson and J.-B. Zuber, *Quantum Field Theory*, McGraw-Hill 1980.
- [4] P. Ramond, *Field Theory: A Modern Primer*, Addison-Wesley 1990.
- [5] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory*, Addison Wesley 1995.
- [6] S. Weinberg, *The Quantum Theory of Fields*, Vols. 1-3, Cambridge 1995
- [7] P. Deligne et al. (editors), *Quantum Fields and Strings: a Course for Mathematicians*, American Mathematical Society 1999.
- [8] A. Zee, *Quantum Field Theory in a Nutshell*, Princeton 2003.
- [9] B. S. DeWitt, *The Global Approach to Quantum Field Theory*, Vols. 1 & 2, Oxford 2003.
- [10] V. P. Nair, *Quantum Field Theory. A Modern Perspective*, Springer 2005.
- [11] T. Banks, *Modern Quantum Field Theory*, Cambridge 2008.
- [12] O. Klein, *Die Reflexion von Elektronen an einem Potentialsprung nach der Relativischen Dynamik von Dirac*, Z. Phys. **53** (1929) 157.
- [13] B. R. Holstein, *Klein's paradox*, Am. J. Phys. **66** (1998) 507.
- [14] N. Dombey and A. Calogeracos, *Seventy years of the Klein paradox*, Phys. Rept. **315** (1999) 41.  
N. Dombey and A. Calogeracos, *History and Physics of the Klein Paradox*, Contemp. Phys. **40** (1999) 313 (quant-ph/9905076).
- [15] F. Sauter, *Zum Kleinschen Paradoxon*, Z. Phys. **73** (1932) 547.
- [16] H. B. G. Casimir, *On the attraction between two perfectly conducting plates*, Proc. Kon. Ned. Akad. Wet. **60** (1948) 793.
- [17] G. Plunien, B. Müller and W. Greiner, *The Casimir Effect*, Phys. Rept. **134** (1986) 87.  
K. A. Milton, *The Casimir Effect: Physical Manifestation of Zero-Point Energy*, (hep-th/9901011).  
K. A. Milton, *The Casimir effect: recent controversies and progress*, J. Phys. **A37** (2004) R209 (hep-th/0406024).  
S. K. Lamoreaux, *The Casimir force: background, experiments, and applications*, Rep. Prog. Phys. **68** (2005) 201.
- [18] M. J. Sparnaay, *Measurement of attractive forces between flat plates*, Physica **24** (1958) 751.
- [19] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover 1972.
- [20] Y. Aharonov and D. Bohm, *Significance of the electromagnetic potentials in the quantum theory*, Phys. Rev. **115** (1955) 485.
- [21] P. A. M. Dirac, *Quantised Singularities in the Electromagnetic Field*, Proc. Roy. Soc. **133** (1931) 60.
- [22] S. Dodelson, *Modern Cosmology*, Academic Press 2003.
- [23] P. A. M. Dirac, *Lectures on Quantum Mechanics*, Dover 2001.



- [24] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems*, Princeton 1992.
- [25] R. Jackiw, *Quantum meaning of classical field theory*, Rev. Mod. Phys. **49** (1977) 681  
R. Jackiw, *Introduction to the Yang-Mills quantum theory*, Rev. Mod. Phys. **52** (1980) 661.
- [26] P. Ramond, *Journeys Beyond the Standard Model*, Perseus Books 1999.  
R. N. Mohapatra, *Unification and Supersymmetry. The Frontiers of Quark-Lepton Physics*, Springer 2003.
- [27] C. P. Burgess, *Goldstone and pseudogoldstone bosons in nuclear, particle and condensed matter physics*, Phys. Rept. **330** (2000) 193 (hep-th/9808176).
- [28] L. Álvarez-Gaumé, *An introduction to anomalies*, in: “Fundamental problems of gauge field theory”, eds. G. Velo and A. S. Wightman, Plenum Press 1986.  
R. A. Bertlmann, *Anomalies in Quantum Field Theory*, Oxford 1996.  
K. Fujikawa and H. Suzuki, *Path Integrals and Quantum Anomalies*, Oxford 2004.  
J. A. Harvey, *TASI lectures on anomalies*, hep-th/0509097.  
L. Álvarez-Gaumé and M. A. Vázquez-Mozo, *Introduction to Anomalies*, Springer (to appear).
- [29] R. Jackiw, *Topological investigations of quantized gauge theories*, in: “Current Algebra and Anomalies”, eds. S. B. Treiman, R. Jackiw, B. Zumino and E. Witten, Princeton 1985.
- [30] S. Adler, *Axial-Vector Vertex in Spinor Electrodynamics*, Phys. Rev. **177** (1969) 2426.  
J. S. Bell and R. Jackiw, *A PCAC puzzle:  $\pi^0 \rightarrow 2\gamma$  in the sigma model*, Nuovo Cimento **A60** (1969) 47.
- [31] J. Steinberger, *On the Use of Subtraction Fields and the Lifetimes of Some Types of Meson Decay*, Phys. Rev. **76** (1949) 1180.
- [32] F. J. Ynduráin, *The Theory of Quark and Gluon Interactions*, Springer 1999.
- [33] G. 't Hooft, *How the instantons solve the U(1) problem*, Phys. Rept. **142** (1986) 357.
- [34] D. G. Sutherland, *Current Algebra and Some Nonstrong Mesonic Decays*, Nucl. Phys. **B2** (1967) 433.  
M. J. G. Veltman, *Theoretical aspects of high-energy neutrino interactions*, Proc. R. Soc. **A301** (1967) 107.
- [35] S. L. Adler and W. A. Bardeen, *Absence of higher order corrections in the anomalous axial vector divergence equation*, Phys. Rev. **182** (1969) 1517.
- [36] E. Witten, *An SU(2) anomaly*, Phys. Lett. **B117** (1982) 324.
- [37] S. Eidelman et al. *Review of Particle Physics*, Phys. Lett. **B592** (2004) 1 (<http://pdg.lbl.gov>).
- [38] D. J. Gross and F. Wilczek, *Ultraviolet behavior of nonabelian gauge theories*, Phys. Rev. Lett. **30** (1973) 1343.
- [39] H. D. Politzer, *Reliable perturbative results for strong interactions?*, Phys. Rev. Lett. **30** (1973) 1346.
- [40] G. 't Hooft, remarks at the *Colloquium on Renormalization of Yang-Mills fields and applications to particle physics*, Marseille 1972.
- [41] I. B. Khriplovich, *Green's functions in theories with a non-abelian gauge group*, Yad. Fiz. **10** (1969) 409 [Sov. J. Nucl. Phys. **10** (1970) 235].  
M. V. Terentiev and V. S. Vanyashin, *The vacuum polarization of a charged vector field*, Zh. Eksp. Teor. Fiz. **48** (1965) 565 [Sov. Phys. JETP **21** (1965) 375].
- [42] K. G. Wilson, *Renormalization group and critical phenomena 1. Renormalization group and the Kadanoff scaling picture*, Phys. Rev. **B4** (1971) 3174.  
K. G. Wilson, *Renormalization group and critical phenomena 2. Phase space cell analysis of critical behavior*, Phys. Rev. **B4** (1971) 3184  
K. G. Wilson, *The renormalization group and critical phenomena*, Rev. Mod. Phys. **55** (1983) 583.
- [43] L. P. Kadanoff, *Scaling Laws for Ising Models Near  $T_c$* , Physics **2** (1966) 263.

- [44] J. Schwinger, *On Gauge Invariance and Vacuum Polarization*, Phys. Rev. **82** (1951) 664.
- [45] E. Brezin and C. Itzykson, *Pair Production in Vacuum by an Alternating Field*, Phys. Rev. **D2** (1970) 1191.
- [46] S. W. Hawking, *Particle Creation by Black Holes*, Commun. Math. Phys. **43** (1975) 199.
- [47] M. K. Parikh and F. Wilczek, *Hawking Radiation as Tunneling*, Phys. Rev. Lett. **85** (2000) 5042 (hep-th/9907001)
- [48] Yu. A. Golfand and E. P. Likhtman, *Extension of the Algebra of Poincaré group generators and violations of P-invariance*, JETP Lett. **13** (1971) 323.  
D. V. Volkov and V. P. Akulov, *Is the Neutrino a Goldstone Particle*, Phys. Lett. **B46** (1973) 109.  
J. Wess and B. Zumino, *A Lagrangian Model Invariant under Supergauge Transformations*, Phys. Lett. **B49** (1974) 52.
- [49] R. Haag, J. Łopuszański and M. Sohnius, *All possible generators of supersymmetries of the S-matrix*, Nucl. Phys. **B88** (1975) 257.

## Basics of QCD for the LHC: $pp \rightarrow H + X$ as a case study

*F. Maltoni*

Centre for Cosmology, Particle Physics and Phenomenology (CP3)  
Université Catholique de Louvain, Louvain-la-Neuve, Belgium

### Abstract

Quantum Chromo Dynamics (QCD) provides the theoretical framework for any study of TeV scale physics at LHC. Being familiar with the basic concepts and techniques of QCD is therefore a must for any high-energy physicist. In these notes we consider Higgs production via gluon fusion as an example on how accurate and flexible predictions can be obtained in perturbative QCD. We start by illustrating how to calculate the total cross section at the leading order (yet one loop) in the strong coupling  $\alpha_S$  and go through the details of the next-to-leading order calculation eventually highlighting the limitations of fixed-order predictions at the parton level. Finally, we briefly discuss how more exclusive (and practical) predictions can be obtained through matching/merging fixed-order results with parton showers.

## 1 Introduction

Strongly interacting particles can be described in terms of a  $SU(3)$  gauge theory field theory involving gluons and quarks:

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4}G^{\mu\nu,a}G_{\mu\nu}^a + \sum_f \bar{\psi}_i^f i\not{D}_{ij} \psi_j^f, \quad (1)$$

where the sum runs over the quark flavors,

$$\begin{aligned} G_{\mu\nu}^a &= \partial_\mu A_\nu^a - \partial_\nu A_\mu^a - g_s f^{abc} A_\mu^b A_\nu^c, \\ D_{\mu,ij} &= \partial_\mu \delta_{ij} + ig_s t_{ij}^a A_\mu^a, \end{aligned}$$

and  $t_{ij}^a$  are the Gell-Mann matrices in the fundamental representation and  $f^{abc}$  are the structure functions of  $SU(3)$ , with

$$[t^a, t^b] = i f^{abc} t^c. \quad (2)$$

Notwithstanding its apparent simplicity, QCD is an amazingly rich theory which is able to account for a wide diversity of phenomena, ranging from really strong (non-perturbative) interactions at low scales, below 1 GeV, to rather weak (perturbative) interactions up to scales of the TeV at colliders, from low density to high density states such as those happening in nuclei collisions or inside stars, from low to high temperatures. For proton-proton collisions at the LHC, where one can consider zero temperature and density, QCD is complicated enough that we have no means available (for the moment!) to solve it exactly and we have to resort to a variety of approximate methods, including perturbation theory (when the coupling is small) and lattice calculations (when the coupling is large). Thanks to the work of theoretical and experimental physicists over the last forty years we are convinced that QCD is a good theory of the strong interactions, of course in the range of energies explored so far and to the level of the theoretical accuracy that can be achieved with current technologies.

There are many excellent references on QCD with applications to collider physics, from books, (e.g., [1]) to review articles, to write-up of lectures given in schools, and in particular some of those given at the CERN schools over the years. My lectures at the school were largely based on the inspiring ones by Michelangelo Mangano [2], Paolo Nason [3] and on the most recent ones by Gavin Salam [4],

which I warmly recommend. In these notes, I'll present a case study, i.e. how QCD can make accurate predictions for Higgs production in gluon fusion at the LHC. The aim is to see the basic concepts at work for a realistic and very important process so to verify their understanding and also to have a closer look at the basic techniques used to perform such calculations. When needed and to avoid repetitions, I will refer to specific sections of Ref. [4] as [QCD: Section number] where the reader will find further information on the basic concepts. Links to simple Mathematica® notebooks with the calculations described below can be found at <http://maltoni.home.cern.ch/>.

## 2 Higgs cross section at the LHC

The factorisation theorem states that the total cross section for the inclusive production of Higgs at the LHC can be written as <sup>1</sup>

$$\sigma(H + X) = \sum_{i,j} \int dx_1 f_i(x_1, \mu_F) \int dx_2 f_j(x_2, \mu_F) \times \hat{\sigma}_{ij \rightarrow H+X}(s, m_H, \mu_F, \mu_R), \quad (3)$$

where the  $f_{i/j}(x, \mu_F)$  are the parton distributions functions (long distance term, non-perturbatively calculable) and  $\hat{\sigma}$  is the partonic cross section (short distance term, calculable in perturbation theory).

$\hat{\sigma}$  can be written as an expansion in  $\alpha_S$ :

$$\begin{aligned} \hat{\sigma}(ij \rightarrow H + X) &= \hat{\sigma}^{(0)}(ij \rightarrow H) \\ &+ \hat{\sigma}^{(1)}(ij \rightarrow H + \text{up to 1 parton}) \\ &+ \hat{\sigma}^{(2)}(ij \rightarrow H + \text{up to 2 partons}) \\ &+ \dots \end{aligned} \quad (4)$$

where the first term gives the leading order (LO) approximation and it is of order  $\alpha_S^2$ , the second next-to-leading (NLO) order ( $\alpha_S^3$ ) and so on.

It is interesting to know how the Higgs predictions improved and evolved over time. The LO production was considered a long ago [5], the next-to-leading order (NLO) QCD corrections [6–9] were calculated decades ago in the so-called effective field theory (HEFT) approximation (which will be explained in the following) as well in the full SM and found to be very large ( $\sigma^{\text{NLO}}/\sigma^{\text{LO}} \sim 2$ ). This motivated the formidable endeavour of the next-to-next-to-leading order (NNLO) QCD calculations, which have been fully evaluated in HEFT [10–12]. Given that corrections to the HEFT been estimated through a power expansion [13–16] and found to have a negligible impact on total rates, NNLO is the current state of the art for fixed-order predictions.

Before going into the details of the computation of the Higgs cross section, let us remind a few general important points that are relevant for any computation in QCD.

- At LO the factorisation theorem reduces to the parton model: the parton distribution functions  $f_i(x)$  are just the probabilities (and therefore positive-definite) of finding a given parton in the initial state hadrons at a given resolution scale  $\mu_F$  and  $\hat{\sigma}$  gives the probability that such partons with a total energy  $s = x_1 x_2 S$  will "fuse" into a Higgs.
- Total cross sections are the first and simplest example of a larger class of observables, called Infrared Safe (IS) quantities [QCD:2.3.2], which can be consistently computed in QCD and then compared to experimental data. Such quantities always need to be (at least to some degree) inclusive on possible extra radiation and in particular resilient under soft and/or collinear radiation. The

---

<sup>1</sup>Be careful here as for simplicity we adopt the usual pragmatic approach on Higgs production at the LHC and imagine it coming from different channels: gluon-gluon fusion, vector-boson-fusion, vector-boson-associated...and so on. We restrict the discussion to the first one which is the leading mechanism. In fact, various channels overlap if contributions are organized as powers of strong and weak couplings (e.g.,  $gg \rightarrow H$  appears at the same order in  $\alpha_S$  and  $y_t$  as  $gg \rightarrow t\bar{t}H$ ) and in general they mix-up once higher-order QCD and EW corrections are included. The separation into channels is anyway useful from the experimental point of view as they typically lead to different final state signatures.

most known example of IS quantities beyond total cross sections are jets [QCD:5]. The constraint of infrared safety becomes non-trivial already at NLO for Eq. (3).

- Total cross sections always inclusive of any possible extra QCD radiation in the event, hereby denoted by  $X$ , even when the calculation is performed at LO. In this case, extra radiation up to the scale  $\mu_F$  is accounted for by the parton distribution function's (PDF), while hard radiation is consistently neglected being of higher order ( $\alpha_S$ ). Alternatively, one can prove that the total cross section for producing "just a Higgs", i.e., Higgs + no resolvable radiation at an arbitrary small scale is exactly zero at all orders in perturbation theory.
- A very important point to always keep in mind is that the the "adjectives" LO, NLO, NNLO need to be always referred to a specific observable, i.e. different observables in a given calculation can be predicted at a different order. For example, when talking about a "NNLO calculation for Higgs production in gluon fusion", what is really meant is that the total inclusive cross section is known at NNLO. The same calculation can predict the rate for Higgs+1 jet (inclusive and exclusive) at NLO and Higgs+2 jets only at LO (where exclusive and inclusive is the same).
- Beyond LO, the separation between long-distance and short-distance physics as described by  $\mu_F$  (and also  $\mu_R$ ) becomes non-trivial.  $\mu_F$  and  $\mu_R$  represent arbitrary scales in the calculation, whose dependence is generated by the truncation of the perturbative expansion at a given order. Exploiting the fact that physical results must be independent on such scales one finds renormalisation-group type equations, such as the  $\beta$  function of QCD [QCD:1.2.3] and the so-called DGLAP evolution equations for the PDF's [QCD:3.2].
- The residual dependence of  $\sigma$  on  $\mu_F$  and  $\mu_R$  at any given order in perturbation theory is often used to gauge the accuracy of the predictions [QCD:4.4.1]. This is by itself a very crude approximation, while the towers of leading (subleading,...) log's of the scales can be predicted at all orders in perturbation theory, only an explicit computation is able to provide the finite terms at higher orders. In practice, it is common to choose central scales as the typical hard scale in a process and vary them independently between 1/2 and 2 to identify an uncertainty. However, no solid and unique procedure exists to identify central reference values and variation intervals and to associate a confidence level. However, milder scale dependence of higher-order results compared to lower ones is always used to gauge the improvement on the accuracy of a given prediction.

### 3 $pp \rightarrow H + X$ at leading order

At LO Eq. 3 can be rewritten as

$$\sigma^{\text{LO}}(H + X) = \int_{\tau_0}^1 dx_1 \int_{\tau_0/x_1}^1 dx_2 f_g(x_1, \mu_F) f_g(x_2, \mu_F) \times \hat{\sigma}^{(0)}(gg \rightarrow H), \quad (5)$$

where  $\tau_0 = m_H^2/S$  and  $s = x_1 x_2 S$ .  $\hat{\sigma}$  for a  $2 \rightarrow 1$  process can be rewritten as

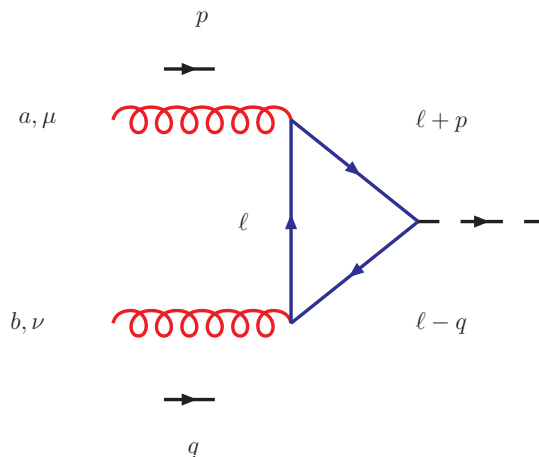
$$\begin{aligned} \hat{\sigma} &= \frac{1}{2s} \overline{|\mathcal{A}|^2} \frac{d^3P}{(2\pi)^3 2E_H} (2\pi)^4 \delta^4(p + q - P_H) \\ &= \frac{1}{2s} \overline{|\mathcal{A}|^2} 2\pi \delta(s - m_H^2), \end{aligned} \quad (6)$$

where

$$\tau \equiv x_1 x_2 = \frac{S}{s}, \quad \tau_0 = \frac{m_H^2}{S}. \quad (7)$$

Performing the change of variables  $x_1, x_2 \rightarrow \tau, y$  with  $x_1 \equiv \sqrt{\tau} e^y$ ,  $x_2 \equiv \sqrt{\tau} e^{-y}$  (verify that the jacobian  $J$  is equal to 1) the change of the integration limits and the result becomes

$$\sigma^{\text{LO}}(H + X) = \frac{\pi \overline{|\mathcal{A}|^2}}{m_H^2 S} \int_{\log \sqrt{\tau_0}}^{-\log \sqrt{\tau_0}} dy x g(\sqrt{\tau_0} e^y) g(\sqrt{\tau_0} e^{-y}). \quad (8)$$



**Fig. 1:** Representative Feynman diagram for the process  $gg \rightarrow H$ . Another diagram, the one with the gluons exchanged, contributes to the total amplitude.

This expression shows that for the cross section of a  $2 \rightarrow 1$  process at LO, the contribution from the parton distributions (a quantity known as gluon-gluon luminosity) factorises from the dynamics ( $|\mathcal{A}|^2$ ). The gluon-gluon luminosity depends only on the kinematics in the limits of integration and can be computed once for all for each Higgs mass. The problem is therefore reduced to the computation of the amplitude  $\mathcal{A}$ .

### 3.1 My first loop (yet finite!) amplitude: $gg \rightarrow H$

Being a color singlet, the Higgs does not couple directly to gluons. However, as no fundamental symmetry forbidding it is present<sup>2</sup> it can via a loop of a colored and massive particle. In the SM such states are the heavy quarks. Let us consider one quark at the time, i.e., the diagram(s) shown in Fig. 1. The first observation to make, even before starting the calculation, is that even though a triangle loop in general can give rise to divergences, both in the ultra-violet (UV) and in the infrared (IR), in this case we expect a finite result. There are several different ways of convincing that this must be the case. A simple one goes as follows. Divergent terms always factorize over lower order amplitudes. The one-loop amplitude is the first non-zero term contributing to  $gg \rightarrow H$  in the perturbative expansion. Therefore there cannot be any divergence. A finite amplitude, however, does not mean that a consistent regularisation procedure is not needed. The reason is that in intermediate steps of the calculation infinities are found that cancel at the end, yet might leave finite terms. As we will see in  $gg \rightarrow H$  such finite terms are actually necessary to guarantee the gauge invariance of the result, clearly showing that there is no ambiguity in the procedure.<sup>3</sup>

To evaluate the diagram of Fig. 1 (there are actually two diagrams, the one shown and another one with the gluons exchanged. They give the same contribution so we'll just multiply our final result by two), we employ use dimensional regularisation in  $d = 4 - 2\epsilon$  dimensions.<sup>4</sup>

<sup>2</sup>In fact, classically, scale invariance would forbid such a coupling. However, scale invariance is broken by renormalisation and therefore it is not a symmetry.

<sup>3</sup>Less obvious is the case of  $\gamma\gamma \rightarrow H$  where the contribution coming from gauge bosons loop has to be done in different gauges (or via low-energy-theorems) to prove the uniqueness and the correctness of dimensional regularisation procedure. Interestingly enough, people seem to forget this fact quite regularly over the years.

<sup>4</sup>Dimensional regularisation comes in several different flavors and attention has to be paid to the details of the implementation. All formulas quoted in the main body of these lecture notes are in the so-called Conventional Dimensional Regularization (CDR) which is the regularisation procedure where the  $\overline{\text{MS}}$  scheme is defined. In practice, NLO calculations nowadays are done in a different scheme which limits the use of the  $d$ -dimensional Dirac algebra to the loop computation.

Using the QCD Feynman rules [QCD: Fig. 3] and the Yukawa interaction, the expression for the amplitude corresponding to the diagram of Fig. 1 reads:

$$i\mathcal{A} = -(-ig_s)^2 \text{Tr}(t^a t^b) \left( \frac{-im_Q}{v} \right) \int \frac{d^d \ell}{(2\pi)^d} \frac{t^{\mu\nu}}{\text{Den}} (i)^3 \epsilon_\mu(p) \epsilon_\nu(q) \quad (9)$$

where the overall minus sign is due to the closed fermion loop.<sup>5</sup> The denominator is  $\text{Den} = (\ell^2 - m_Q^2)[(\ell + p)^2 - m_Q^2][(\ell - q)^2 - m_Q^2]$ . Employing the usual Feynman parametrization method to combine the denominators of the loop integral into one:

$$\frac{1}{ABC} = 2 \int_0^1 dx \int_0^{1-x} dy \frac{1}{[Ax + By + C(1-x-y)]^3} \quad (10)$$

one obtains

$$\frac{1}{\text{Den}} = 2 \int dx dy \frac{1}{[\ell^2 - m_Q^2 + 2\ell \cdot (px - qy)]^3}. \quad (11)$$

The next step is to shift the integration momenta to  $\ell' = \ell + px - qy$  so the denominator takes the form

$$\frac{1}{\text{Den}} \rightarrow 2 \int dx dy \frac{1}{[\ell'^2 - m_Q^2 + m_H^2 xy]^3}. \quad (12)$$

The numerator of the loop integral in the shifted loop momentum becomes

$$\begin{aligned} t^{\mu\nu} &= \text{Tr} \left[ (\ell + m_Q) \gamma^\mu (\ell + \not{p} + m_Q) (\ell - \not{q} + m_Q) \gamma^\nu \right] \\ &= 4m_Q \left[ g^{\mu\nu} (m_Q^2 - \ell^2 - \frac{m_H^2}{2}) + 4\ell^\mu \ell^\nu + p^\nu q^\mu \right]. \end{aligned} \quad (13)$$

where we have used the fact that for transverse gluons,  $\epsilon(p) \cdot p = 0$  and so terms proportional to the external momenta,  $p_\mu$  or  $q_\nu$ , have been dropped. The above expression shows already several interesting aspects.

The first one is that the trace is proportional to the heavy quark mass. This can be easily understood as an effect of the spin-flip coupling of the Higgs. Gluons or photons do not change the spin of the fermion, as vectors map left (right) spinors into left (right) spinors, while the scalars do couple left (right) spinors with right (left) ones. If the quark circulating in the loop is massless then the trace vanishes due to helicity conservation, independently of the actual Yukawa coupling. This is the reason why even when the Yukawa coupling of the light quark and the Higgs is enhanced (such as in SUSY or 2HDM with large  $\tan \beta$ ), the contribution is anyway suppressed by the kinematical mass.

The second point is that simple power counting shows that the terms proportional to the squared loop momentum  $\ell^2$  and  $\ell^\mu \ell^\nu$  give rise to UV divergences. This means that an intermediate and consistent regularisation prescription is needed for intermediate manipulations and that divergences will have to cancel in the final result.

By shifting momenta in the numerator, dropping terms linear in  $\ell'$  and using the relation

$$\int d^d k \frac{k^\mu k^\nu}{(k^2 - C)^m} = \frac{1}{d} g^{\mu\nu} \int d^d k \frac{k^2}{(k^2 - C)^m} \quad (14)$$

to write the amplitude in the form

$$i\mathcal{A} = -\frac{2g_s^2 m_Q^2}{v} \delta^{ab} \int \frac{d^d \ell'}{(2\pi)^d} \int dx dy \left\{ g^{\mu\nu} \left[ m_Q^2 + \ell'^2 \left( \frac{4-d}{d} \right) + m_H^2 \left( xy - \frac{1}{2} \right) \right] \right\}$$

<sup>5</sup> $\epsilon_\mu(p)$  are the transverse gluon polarizations.

$$+p^\nu q^\mu (1 - 4xy) \left. \vphantom{\frac{2dxdy}{(\ell^2 - m_Q^2 + m_H^2 xy)^3}} \right\} \frac{2dxdy}{(\ell^2 - m_Q^2 + m_H^2 xy)^3} \epsilon_\mu(p) \epsilon_\nu(q). \quad (15)$$

This expression shows that if one computes the integral in  $d = 4$ , the UV divergent term is absent. For  $d = 4 - 2\epsilon$ , however, this gives rise to a left-over finite piece, as the scalar integrals are given by

$$\begin{aligned} \int \frac{d^d \ell}{(2\pi)^d} \frac{\ell^2}{(\ell^2 - C)^3} &= \frac{i}{32\pi^2} (4\pi)^\epsilon \frac{\Gamma(1 + \epsilon)}{\epsilon} (2 - \epsilon) C^{-\epsilon} \\ \int \frac{d^d \ell}{(2\pi)^d} \frac{1}{(\ell^2 - C)^3} &= -\frac{i}{32\pi^2} (4\pi)^\epsilon \Gamma(1 + \epsilon) C^{-1-\epsilon}. \end{aligned} \quad (16)$$

So it is manifest that the divergence  $1/\epsilon$  cancels against the  $(4 - d)/d$  term leaving a finite piece, which in fact ensures that the final result is gauge invariant. By combining it with the other terms in the squared parenthesis we obtain

$$\mathcal{A}(gg \rightarrow H) = -\frac{\alpha_S m_Q^2}{\pi v} \delta^{ab} \left( g^{\mu\nu} \frac{m_H^2}{2} - p^\nu q^\mu \right) \epsilon_\mu(p) \epsilon_\nu(q) \int dxdy \left( \frac{1 - 4xy}{m_Q^2 - m_H^2 xy} \right). \quad (17)$$

(Note that we have multiplied by 2 in Eq. (17) to include the diagram where the gluon legs are crossed.) The Feynman integral of Eq. (17) can easily be performed to find an analytic result if desired. Note that the tensor structure could have been predicted from the start by imposing gauge invariance, i.e.,  $p^\mu \mathcal{A}^{\mu\nu} = q^\nu \mathcal{A}^{\mu\nu} = 0$ . By defining  $I(a)$  as

$$I(a) \equiv \int_0^1 dx \int_0^{1-x} dy \frac{1 - 4xy}{1 - axy}, \quad a = \frac{m_H^2}{m_Q^2}, \quad (18)$$

one can factorise a  $1/m_Q^2$  out of the integral and cancel the overall  $m_Q^2$  in front of the amplitude (17). In other terms the heavy quark mass dependence is confined in  $I(a)$ .

For a light quark,  $m_Q \ll m_H$ ,

$$I(a) \xrightarrow{a \rightarrow \infty} -\frac{1}{2a} \log^2 a = -\frac{m_Q^2}{2m_H^2} \log^2 \frac{m_Q^2}{m_H^2}, \quad (19)$$

showing that in the Standard Model the charm and bottom quark contributions are strongly suppressed by the square of the quark mass over Higgs mass ratio and come with a minus sign (with respect to the top-quark one).

The opposite limit,  $m_H \ll m_Q$ ,

$$I(a) \xrightarrow{a \rightarrow 0} \frac{1}{3}, \quad (20)$$

which is found to be an extremely good approximation even for  $m_Q \sim m_H$ , is quite surprising at first. In this case the amplitude reads

$$\mathcal{A}(gg \rightarrow H) \xrightarrow{m_Q \gg m_H} -\frac{\alpha_S}{3\pi v} \delta^{ab} \left( g^{\mu\nu} \frac{m_H^2}{2} - p^\nu q^\mu \right) \epsilon_\mu(p) \epsilon_\nu(q). \quad (21)$$

i.e., the amplitude  $gg \rightarrow H$  becomes *independent* of the mass of the heavy fermion in the loop. This is a special case of a general low energy theorem (which holds in the  $p_H \rightarrow 0$  limit) that states that if the colored particle mass, independently of the other quantum numbers such as its spin acquires (all of) its mass via the Higgs mechanism, it will contribute to the amplitude  $gg \rightarrow H$  independently of its mass. In other words  $gg \rightarrow H$  acts as a counter of heavy colored particles. In a four generation scenario, for instance, the contribution from the  $t'$  and  $b'$  would lead to a factor of three increase at the amplitude level, i.e. a factor 9 at the cross section level. Note that this is in an apparent contradiction with of our intuition



that heavy particles should decouple and not affect the physics at lower energy. The heavy states would not decouple because of our assumption that their (whole) mass is due to electroweak symmetry breaking and the interaction with the Higgs. Another interesting case is that of SUSY, where down-type and up-type quarks can couple differently to the Higgs(es) and other colored states (squarks) are present in the spectrum. At large  $\tan \beta$ , i.e. when  $m_b \tan \beta \simeq m_t$ , the Higgs bottom couplings are enhanced by a factor  $\tan \beta$ , while those of the top suppressed by a  $\cot \beta$ . However, the scaling with masses is different in the two limits and the contribution from the bottom anyway suppressed by  $m_Q/m_H$ . In addition, the the two contributions will have an opposite sign so that will actually interfere destructively in the amplitude squared. What about the squark contributions? Being heavy scalars and therefore coming with an opposite sign shouldn't the stop cancel exactly the contributions from the top and the others squarks give the dominant contribution? In this case, one has to remember that in (possibly) realistic SUSY models the mass of a squark has two sources: one from the coupling to the Higgs vev, which due to SUSY, it is exactly equal to the SM partner coupling and the other from the SUSY soft-breaking terms. For light quarks the latter are by far dominant giving a scaling for  $\mathcal{A}$  of the type  $m_q/m_{\tilde{q}}$ , so highly suppressed and decoupling. A light stop instead,  $m_{\tilde{t}} \simeq m_t$  could lead to a possibly strong suppression of  $\mathcal{A}$ .

### 3.2 Total cross section at the LHC at LO

The result can be written as:

$$\sigma^{\text{LO}}(pp \rightarrow H + X) = \frac{\alpha_S^2(\mu_R)}{64\pi v^2} \left| I\left(\frac{m_H^2}{m_Q^2}\right) \right|^2 \tau_0 \int_{\log \sqrt{\tau_0}}^{-\log \sqrt{\tau_0}} dy g(\sqrt{\tau_0} e^y, \mu_F) g(\sqrt{\tau_0} e^{-y}, \mu_F) \quad (22)$$

Using LO PDF's available in public libraries, such as LHAPDF [17] one can easily compute the gluon-gluon luminosity and therefore the LO Higgs cross section at the LHC14, see Fig. 2. An example is given in a Mathematica® notebook that can be found at the web address mentioned at the end of the Introduction. An interesting exercise is to vary the value of the renormalisation and factorisation scales around the natural central choice  $\mu_R = \mu_F = m_H$  to try to estimate the unknown higher-orders terms in the perturbative expansion. It has to be noted that at LO, the cross section depends on  $\mu_R$  only through  $\alpha_S(\mu_R)$  which appears in the short distance coefficient and therefore as an overall factor  $\alpha_S^2$ , and depends on  $\mu_F$  only via the PDF's (both dependences are of logarithmic nature, as the application of the renormalisation group equations easily shows). In other words the dependence on the scales is maximal as there is no explicit dependence on the log of the scales in the short distance coefficients that can compensate those in the coupling and in the PDF's. At this order, this is consistent as scale changes correspond to a change of at least one order in  $\alpha_S$  more and in a LO computation only the first term in the perturbative expansion is present. The result of varying the scales independently  $1/2 m_H < \mu_R, \mu_F < 2 m_H$  with  $1/2 < \mu_F/\mu_R < 2$  in the LO predictions for the LHC is shown in Fig. 9 for different Higgs masses. Result are normalized to the central reference choice  $\mu_R = \mu_F = m_H$ .

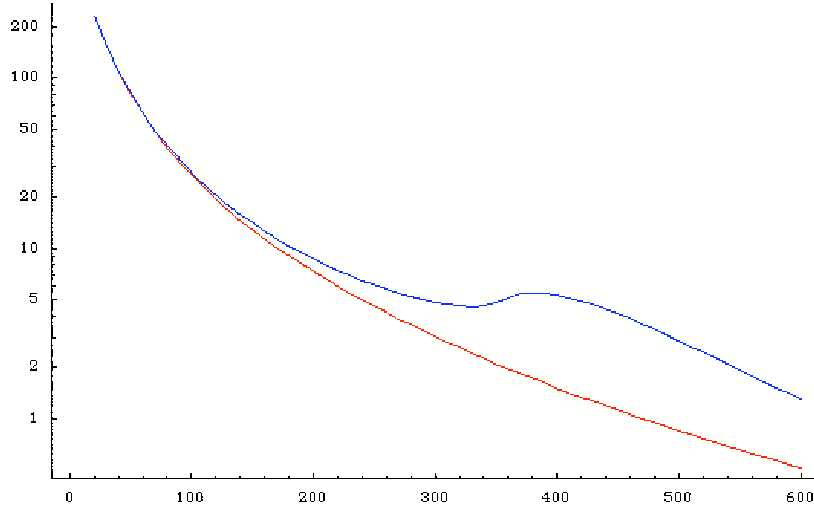
## 4 Higgs Effective field theory

The main result of the simple calculation  $gg \rightarrow H$  is that gluon fusion is basically independent of the heavy quark mass for a light Higgs boson. The result of Eq. (33) can be easily derived starting from the effective vertex,

$$\begin{aligned} \mathcal{L}_{\text{eff}} &= \frac{\alpha_S}{12\pi} G_{\mu\nu}^a G^{a\mu\nu} \left(\frac{H}{v}\right) \\ &= \frac{\beta_F}{g_s} G_{\mu\nu}^a G^{a\mu\nu} \left(\frac{H}{2v}\right) (1 - \delta), \end{aligned}$$

where

$$\beta_F = \frac{g_s^3 N_F}{24\pi^2} \quad (23)$$



**Fig. 2:** Example of a plot for the LO cross section for  $pp \rightarrow H$  at the LHC14 (pb) as a function of the Higgs mass (GeV) obtained with Mathematica® notebook available from the author (link in the text). The red (lower) curve is the large top-mass limit, while the blue (upper) curve is the result with full top-mass dependence.

is the contribution of heavy fermion loops to the  $SU(3)$  beta function and  $\delta = 2\alpha_S/\pi$ .<sup>6</sup> ( $N_F$  is the number of heavy fermions with  $m \gg m_H$ .) The effective Lagrangian of Eq. (23) gives  $ggH$ ,  $gggH$  and  $ggggH$  vertices and can be used to compute the radiative corrections of  $\mathcal{O}(\alpha_S^3)$  to gluon production. The correction in principle involves 2-loop diagrams. However, using the effective vertices from Eq. (23), the  $\mathcal{O}(\alpha_S^3)$  corrections can be found from a 1-loop calculation. To fix the notation we shall use

$$\mathcal{L}_{\text{eff}} = -\frac{1}{4}AHG_{\mu\nu}^a G^{a,\mu\nu}, \quad (24)$$

where  $G_{\mu\nu}^a$  is the field strength of the  $SU(3)$  color gluon field and  $H$  is the Higgs-boson field. The effective coupling  $A$  is given by

$$A = \frac{\alpha_S}{3\pi v} \left( 1 + \frac{11}{4} \frac{\alpha_S}{\pi} \right), \quad (25)$$

where  $v$  is the vacuum expectation value parameter,  $v^2 = (G_F\sqrt{2})^{-1} = (246)^2 \text{ GeV}^2$  and the  $\alpha_S$  correction is included, as discussed above. The effective Lagrangian generates vertices involving the Higgs boson and two, three or four gluons. The associated Feynman rules are displayed in Fig. 3. The two-gluon–Higgs-boson vertex is proportional to the tensor

$$H^{\mu\nu}(p_1, p_2) = g^{\mu\nu} p_1 \cdot p_2 - p_1^\nu p_2^\mu, \quad (26)$$

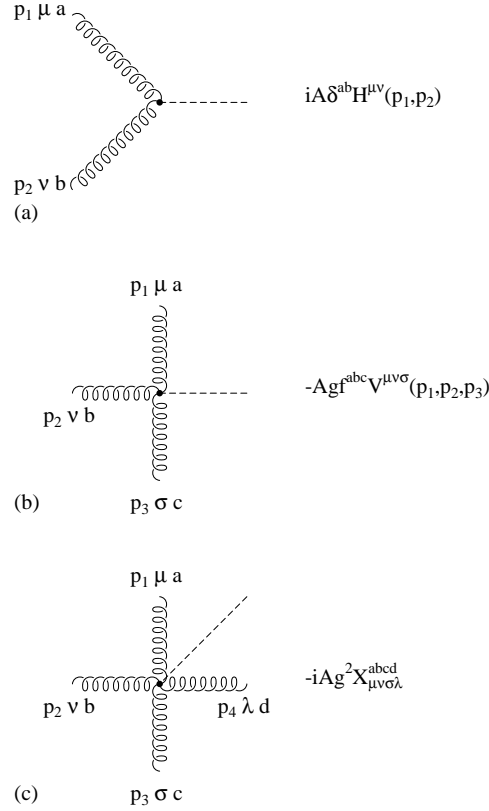
while the vertices involving three and four gluons and the Higgs boson are exactly proportional to their counterparts from pure QCD

$$V^{\mu\nu\rho}(p_1, p_2, p_3) = (p_1 - p_2)^\rho g^{\mu\nu} + (p_2 - p_3)^\mu g^{\nu\rho} + (p_3 - p_1)^\nu g^{\rho\mu}, \quad (27)$$

and

$$X_{abcd}^{\mu\nu\rho\sigma} = f_{abe}f_{cde}(g^{\mu\rho}g^{\nu\sigma} - g^{\mu\sigma}g^{\nu\rho}) + f_{ace}f_{bde}(g^{\mu\nu}g^{\rho\sigma} - g^{\mu\sigma}g^{\nu\rho})$$

<sup>6</sup>The  $(1 - \delta)$  term arises from a subtlety in the use of the low energy theorem. Since the Higgs coupling to the heavy fermions is  $M_f(1 + \frac{H}{v})\bar{f}f$ , the counterterm for the Higgs Yukawa coupling is fixed in terms of the renormalisation of the fermion mass and wavefunction. The beta function, on the other hand, is evaluated at  $q^2 = 0$ . The  $1 - \delta$  term corrects for this mismatch.



**Fig. 3:** Feynman rules in the EFT where the top quark is integrated out. Gluon momenta are outgoing.

$$+ f_{ade} f_{bce} (g^{\mu\nu} g^{\rho\sigma} - g^{\mu\rho} g^{\nu\sigma}). \quad (28)$$

## 5 $gg \rightarrow \text{Higgs}$ @ NLO

The HEFT is clearly a very powerful approximation as it turns a loop computation into a tree-level one. That means that within the HEFT the calculation of the total cross section for Higgs production at NLO will appear as a usual NLO calculation, i.e., involving only one-loop and tree-level diagrams. This is what we describe in this section.

### 5.1 The NLO computation in a nutshell

At NLO Eq. 3 can be rewritten as

$$\begin{aligned} \sigma^{\text{NLO}}(H + X) &= \int_{\tau_0}^1 dx_1 \int_{\tau_0/x_1}^1 dx_2 f_g(x_1, \mu_F) f_g(x_2, \mu_F) [\hat{\sigma}_B^{(0)}(gg \rightarrow H) + \hat{\sigma}_V^{(1)}(gg \rightarrow H)] \\ &+ \sum_{ijk} \int_{\tau_0}^1 dx_1 \int_{\tau_0/x_1}^1 dx_2 f_i(x_1, \mu_F) f_j(x_2, \mu_F) \times \hat{\sigma}_R^{(1)}(ij \rightarrow H k), \end{aligned} \quad (29)$$

where  $\hat{\sigma}^{(0)}(gg \rightarrow H)$  and  $\hat{\sigma}_V^{(1)}(gg \rightarrow H)$  denote the Born-level and the virtual cross sections, while  $\hat{\sigma}_R^{(1)}(ij \rightarrow H k)$  is the real-emission cross section:

$$\hat{\sigma}_{B,V}^{(0,1)}(gg \rightarrow H) = \frac{1}{2s} \overline{|\mathcal{A}_{B,V}|^2} d\Phi_B,$$

$$\hat{\sigma}_R^{(1)}(ij \rightarrow Hk) = \frac{1}{2s} \overline{|\mathcal{A}_R|^2} d\Phi_R,$$

In general, the virtual term contains ultraviolet (UV), soft and collinear divergences. The UV divergences are absorbed by a universal redefinition of the couplings entering at the Born amplitude, as dictated by the renormalisation of the SM. When integrated over the full real phase space, the real term generates soft and collinear divergences, too, and only when *infrared(IR)-safe* quantities are computed, these divergences cancel to yield a finite result. IR-safe observables  $O(\Phi)$  can be best understood by considering the soft or collinear limit in the real phase space, i.e. when the additional parton has low energy or is parallel to another parton. In this limit, an IR-safe observable yields  $\lim O(\Phi_R) = O(\Phi_B)$ , where the Born-level configuration  $\Phi_B$  is obtained from  $\Phi_R$  by eliminating the soft particle (in case of soft singularities) or by merging the collinear particles (in case of collinear singularities).

There several ways to handle the cancellation of the singularities, which fall into two large categories, process-dependent and process-independent methods. In the former, one treats each calculation/process independently and performs manipulations of the integrals over the phase space so to obtain analytic or semi-analytic results.

Process independent methods, on the other hand, are based on a very fundamental result, i.e., that the pattern of the soft and collinear divergences is universal and depends only on the quantum numbers of the initial and final state particles in the Born process. That means that given the Born amplitude, one can predict the divergences that will show up in the virtual contributions and will be then cancelled over integration of the extra radiation in the reals. More importantly, such divergences come in just a handful of different types that can be dealt with once and for all.

Let us now rewrite Eq. (29) in a general and short-hand notation

$$\sigma^{\text{NLO}} \equiv \int d\Phi_B [B(\Phi_B) + V(\Phi_B)] O(\Phi_B) + \int d\Phi_R R(\Phi_R) O(\Phi_R) \quad (30)$$

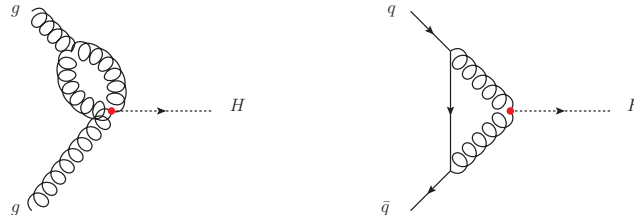
which will be useful in the following. A NLO cross section is written in terms of matrix elements for the Born and virtual integrated over the Born phase space plus the real matrix elements integrated over the real phase space. Within a subtraction method, the real phase space is parametrized in terms of an underlying Born phase space  $\Phi_B$  and a radiation phase space  $\Phi_{R|B}$ . A necessary requirement upon this parametrization is that, in the singular limits, by merging collinear partons, or eliminating the soft parton, the real phase becomes equal to the underlying Born one. Then the expectation value of an IR-safe observable reads

$$\begin{aligned} \int d\sigma^{(\text{NLO})} O(\Phi) &= \int d\Phi_B \left[ B(\Phi_B) + V(\Phi_B) + \int d\Phi_{R|B} S(\Phi_R) \right] O(\Phi_B) \\ &+ \int d\Phi_R [R(\Phi_R) O(\Phi_R) - S(\Phi_R) O(\Phi_B)] . \end{aligned} \quad (31)$$

The third member of the above equation is obtained by adding and subtracting the same quantity from the two terms of the second member. The terms  $S(\Phi_{R|B})$  are the subtraction terms, which contain all soft and collinear singularities of the real-emission term. Using the universality of soft and collinear divergences, they are written in a factorised form as

$$S(\Phi_R) = B(\Phi_B) \otimes \tilde{S}(\Phi_{R|B}), \quad (32)$$

where the  $\tilde{S}(\Phi_{R|B})$  can be composed from universal, process-independent subtraction kernels with analytically known (divergent) integrals. These integral, when summed and added to the virtual term, yield a finite result. The second term of the last member of Eq. (31) is also finite if  $O$  is an IR-safe observable, since by construction  $S$  cancels all singularities in  $R$  in the soft and collinear regions. The most popular subtraction schemes currently used in public NLO codes are based on the dipole subtraction [18] and the



**Fig. 4:** Example of Feynman diagrams giving null contributions to  $ij \rightarrow H$  at one-loop in the HEFT. Bubbles on the gluon legs are zero in dimensional regularisation.  $q\bar{q} \rightarrow H$  is zero at all orders in perturbation theory if  $m_q = 0$  due to chiral symmetry.

so-called FKS scheme [19]. The case of  $gg \rightarrow H$  at NLO is particularly simple as the Born amplitude is a  $2 \rightarrow 1$  process. This means that the integration over phase space of the real corrections is particularly simple and can therefore be done analytically. This has also the pedagogical advantage that shows explicitly where the divergences come from and to “see” the cancellations term by term. We study the process  $gg \rightarrow H$  at NLO, in the large top-quark mass limit. All results given below are in Conventional Dimensional Regularization (CDR), where matrix elements are calculated in  $d$  dimensions, including the Born and real contributions, as well as the integration over phase space [6].

## 5.2 $gg \rightarrow H$ : Born in $d$ dimensions

The Born amplitude is calculated via the HEFT feynman rules. The only difference with respect to the previous calculation stems from the fact that now the computation has to be done in  $d = 4 - 2\epsilon$ -dimensions, with  $\epsilon$  infinitesimal. The phase space do not bring any extra  $\epsilon$  term. However, the matrix element changes

$$\left( g^{\mu\nu} \frac{m_H^2}{2} - p^\nu q^\mu \right)^2 = \frac{1}{4} (d-2) m_H^4, \quad (33)$$

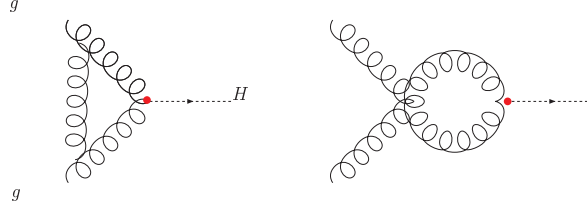
as well as the average over the initial state gluon polarizations which in  $d$ -dimensions are  $d-2$ . This gives

$$\begin{aligned} \hat{\sigma}_B &= \frac{\alpha_S^2}{\pi} \frac{m_H^2}{576v^2s} \frac{\mu^{2\epsilon}}{(1-\epsilon)} \delta(1-z) \\ &\equiv \hat{\sigma}_0 \delta(1-z), \end{aligned} \quad (34)$$

where  $z \equiv m_H^2/s$  is the inelasticity of the process, i.e. the fraction of the parton parton energy that goes into the Higgs (for the Born  $z = 1$ ).  $\mu$  is the usual arbitrary scale that needs to be introduced in dimensional regularisation to correct for the different dimensions and keep the action adimensional ( $\hbar = c = 1$ ). Note that a cross section in  $d$  dimensions has dimensions  $[\sigma] = M^{2-d}$ . Also note that we have defined  $\hat{\sigma}_0$  as containing an explicit factor  $z$ .

## 5.3 $gg \rightarrow H$ : virtual corrections

There are several diagrams appearing at one-loop. Diagrams involving bubbles on the external gluon legs (with 3-point gluon-gluon-gluon and gluon-gluon-Higgs vertexes) give rise to scaleless integrals that are zero in dimensional regularisation, see Fig. 4, left diagram. The  $q\bar{q} \rightarrow H$  process, see Fig 4 right, is



**Fig. 5:** Feynman diagrams giving non-zero contributions to  $gg \rightarrow H$  at one-loop in the HEFT.

proportional to the  $m_q$  parton mass which are taken massless and therefore null at all orders. As a result, only two diagrams are non-zero, i.e., the vertex correction and the bubble with the four gluon vertex as shown in Fig. 5

$$\hat{\sigma}_{\text{tri}} = \hat{\sigma}_0 \delta(1-z) \left[ 1 + \frac{\alpha_S}{2\pi} C_A \left( \frac{\mu^2}{m_H^2} \right)^\epsilon c_\Gamma \left( -\frac{2}{\epsilon^2} + \frac{10}{3\epsilon} + \frac{179}{36} + \pi^2 \right) \right], \quad (35)$$

$$\hat{\sigma}_{\text{bub}} = \hat{\sigma}_0 \delta(1-z) \left[ 1 + \frac{\alpha_S}{2\pi} C_A \left( \frac{\mu^2}{m_H^2} \right)^\epsilon c_\Gamma \left( -\frac{10}{3\epsilon} - \frac{179}{36} \right) \right], \quad (36)$$

where

$$c_\Gamma = (4\pi)^\epsilon \frac{\Gamma(1+\epsilon)\Gamma(1-\epsilon)^2}{\Gamma(1-2\epsilon)}. \quad (37)$$

To obtain the results above, one has to write down the loop amplitudes, perform a few simplifications and the decomposition of the tensor integrals appearing in the amplitudes so to express the results in terms of the following two scalar integrals:

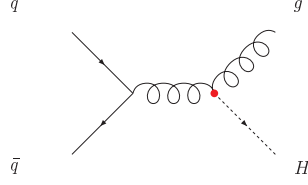
$$\begin{aligned} \mu^{2\epsilon} \int \frac{d^d \ell}{(2\pi)^d} \frac{1}{\ell^2 (\ell + p_H)^2} &= c_\Gamma \left( \frac{\mu^2}{m_H^2} \right)^\epsilon \left( \frac{1}{\epsilon} + 2 \right), \\ \mu^{2\epsilon} \int \frac{d^d \ell}{(2\pi)^d} \frac{1}{\ell^2 (\ell + p_1)^2 (\ell + p_2)^2} &= \frac{c_\Gamma}{2m_H^2} \left( \frac{\mu^2}{m_H^2} \right)^\epsilon \left( \frac{2}{\epsilon^2} - \pi^2 \right), \end{aligned} \quad (38)$$

with  $p_H = p_1 + p_2$ . Summing the contributions of the two diagrams above with the  $\alpha_S$  correction from Eq. (25), we obtain

$$\hat{\sigma}_V = \hat{\sigma}_0 \delta(1-z) \left[ 1 + \frac{\alpha_S}{2\pi} C_A \left( \frac{\mu^2}{m_H^2} \right)^\epsilon c_\Gamma \left( -\frac{2}{\epsilon^2} + \frac{11}{3} + \pi^2 \right) \right], \quad (39)$$

i.e., the total virtual contribution is proportional to the Born amplitude and it contains pole(s) in powers of  $1/\epsilon$ . The fact that the full virtual amplitude is proportional to the Born is due to the simplicity of a  $2 \rightarrow 1$  process. However, in general one can prove that the divergent contributions must be proportional to the Born in the case of collinear (and collinear-soft, the double pole) divergences and to the so-called color-connected Born for the soft ones. Given that the Born amplitude is proportional to  $\alpha_S^2$  and we are calculating QCD corrections, we also expect UV divergences, which are proportional to  $1/\epsilon$ . The fact that apparently we do not see any pole in  $1/\epsilon$  in the result above, it simply means that there is an accidental cancellation between simple poles of IR origin and that of UV origin, as we did not keep them distinct in the calculation. To leave only IR poles in the amplitude to be cancelled with those coming from the real contribution, we therefore proceed here to renormalisation of  $\alpha_S$ . This can be attained by the substitution in  $\hat{\sigma}_0$ , see also [QCD:1.2.3],

$$\alpha_S \rightarrow \alpha_S^{\overline{\text{MS}}}(\mu_R) = \alpha_S \left[ 1 - \frac{\alpha_S}{2\pi} c_\Gamma \left( \frac{\mu^2}{\mu_R^2} \right)^\epsilon \frac{b_0}{\epsilon} \right], \quad (40)$$



**Fig. 6:** Feynman diagrams giving  $q\bar{q}$  real contributions in the infinite top-quark mass limit. These contributions are finite.

where  $b_0 = 11/6 C_A - 2n_f T_F/3$ . The UV-renormalized virtual amplitude is

$$\hat{\sigma}_V^{\overline{\text{MS}}}(gg) = \hat{\sigma}_0 \delta(1-z) \left[ 1 + \frac{\alpha_S}{2\pi} C_A \left( \frac{\mu^2}{m_H^2} \right)^\epsilon c_\Gamma \left( -\frac{2}{\epsilon^2} - \frac{2}{\epsilon} \frac{b_0}{C_A} - 2 \frac{b_0}{C_A} \log \frac{m_H^2}{\mu_R^2} + \frac{11}{3} + \pi^2 \right) \right]. \quad (41)$$

where now the poles in  $1/\epsilon^2$ ,  $1/\epsilon$  are only of IR nature. Another important feature which is manifest in the expression above is the appearance of an explicit log of the renormalisation scale in the short distance part. As mentioned before, this is the improvement expected on the scale dependence of a NLO result: the  $\mu_R$  dependence of the  $\alpha_S^2(\mu_R)$  overall coefficient is exactly cancelled by the explicit log up to order  $\alpha_S^3$ .

## 5.4 Real Contributions

Real corrections imply the calculation of  $2 \rightarrow 2$  tree-level amplitudes and their integration over phase space in  $d$  dimensions. All possible initial and final state partons, gluons, quarks and anti-quarks need to be included,

1.  $q\bar{q} \rightarrow Hg$  + crossing (i.e.,  $\bar{q}q \rightarrow Hg$ ),
2.  $qg \rightarrow Hq$  + crossings (i.e.,  $\bar{q}g \rightarrow H\bar{q}$ ,  $gq \rightarrow Hq$ ,  $g\bar{q} \rightarrow H\bar{q}$ ),
3.  $gg \rightarrow Hg$ .

It is easy to predict which divergences to expect from each of the subprocesses above. The reason is that out of the possible (by Lorentz and color invariance) underlying Born amplitudes, i.e.,  $q\bar{q} \rightarrow H$  and  $gg \rightarrow H$ , the only non-zero one is  $gg \rightarrow H$ . Therefore the first processes must give a finite result when integrated over phase space, the second ones can only contain collinear divergences to be absorbed in quark PDF's, while the last is expected to give rise to soft and collinear divergences, part of which will be absorbed in the gluon PDF's and the rest canceled against those coming from the virtual contributions, Eq. (41).

### 5.4.1 $q\bar{q} \rightarrow Hg$

This contribution, shown in Fig. 6 is finite and can be calculated directly in four dimensions. A simple calculation gives

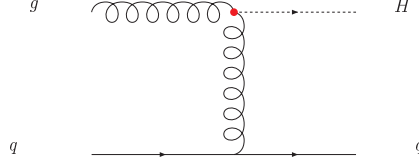
$$\overline{|\mathcal{M}|^2} = \frac{4}{81} \frac{\alpha_S^3}{\pi v^2} \frac{(u^2 + t^2)}{s}, \quad (42)$$

to be integrated over the 4-dimensional phase space

$$d\Phi_2 = \frac{1}{8\pi} (1-z) dv, \quad (43)$$

where  $v = 1/2(1 + \cos \theta)$  and  $z = m_H^2/s$  as usual. Using

$$t = -s(1-z)(1-v), \quad (44)$$



**Fig. 7:** Feynman diagrams giving  $qg$  real contributions in the infinite top-quark mass limit.

$$u = -s(1-z)v, \quad (45)$$

gives

$$\hat{\sigma}_R(q\bar{q}) = \hat{\sigma}_0 \frac{\alpha_S}{2\pi} \frac{64}{27} \frac{(1-z)^3}{z}. \quad (46)$$

#### 5.4.2 $gq \rightarrow Hq$

Let us consider now the contribution from the diagrams with an initial quark, i.e., the process  $gq \rightarrow Hq$ . The  $d$ -dimensional averaged/summed over initial/final state polarizations and colors amplitude is

$$\overline{|\mathcal{M}|^2} = -\frac{1}{54(1-\epsilon)} \frac{\alpha_S^3}{\pi v^2} \frac{(u^2 + s^2) - \epsilon(u+s)^2}{t}. \quad (47)$$

Integrating it over the  $d$ -dimensional phase space

$$d\Phi_2 = \frac{1}{8\pi} \left( \frac{4\pi}{s} \right)^\epsilon \frac{1}{\Gamma(1-\epsilon)} z^\epsilon (1-z)^{1-2\epsilon} v^{-\epsilon} (1-v)^{-\epsilon} dv \quad (48)$$

one gets

$$\hat{\sigma}_R(gq) = \hat{\sigma}_0 \frac{\alpha_S}{2\pi} C_F \left( \frac{\mu^2}{m_H^2} \right)^\epsilon c_\Gamma \left[ -\frac{1}{\epsilon} p_{gq}(z) + z - \frac{3}{2} \frac{(1-z)^2}{z} + p_{gq}(z) \log \frac{(1-z)^2}{z} \right], \quad (49)$$

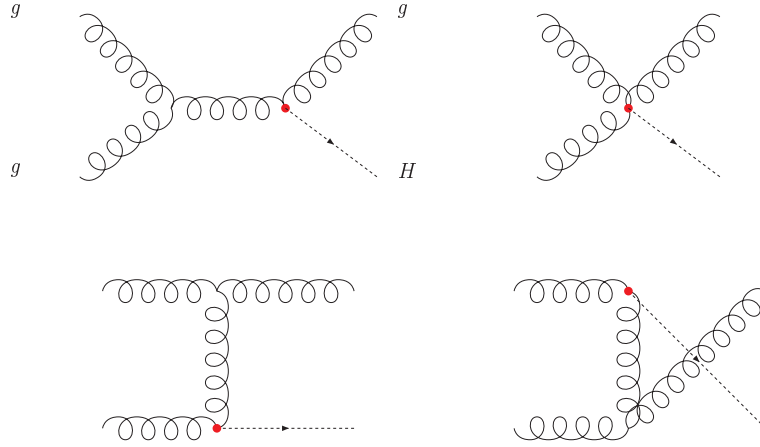
where the  $p_{gq}(z)$  color-stripped Altarelli-Parisi splitting function is given in the Appendix, Eqs. (67). We perform the factorisation of the collinear divergences adding the counterterm

$$\sigma_{\text{c.t.}}^{\text{coll.}}(gq) = \sigma_0 \frac{\alpha_S}{2\pi} \left[ \left( \frac{\mu^2}{\mu_F^2} \right)^\epsilon \frac{c_\Gamma}{\epsilon} P_{gq}(z) \right]. \quad (50)$$

We note that in fact in CDR the cross section factorises over the  $d$ -dimensional splitting functions Eqs. (68). However, the collinear counter-term in  $\overline{\text{MS}}$  is defined with the 4-dimensional Altarelli-Parisi splitting functions, Eqs. (67), and that is why we have written the result above in terms of  $p_{gq}(z)$  leaving out a finite term  $z$  (also note that our definition of  $\sigma_0$ , Eq. (34), contains a factor  $z$ ). This gives

$$\begin{aligned} \hat{\sigma}_R^{\overline{\text{MS}}}(gq) &= \hat{\sigma}_R(gq) + \hat{\sigma}_{\text{c.t.}}^{\text{coll.}}(gq) \\ &= \sigma_0 \frac{\alpha_S}{2\pi} C_F \left[ p_{gq}(z) \log \frac{m_H^2}{\mu_F^2} + p_{gq}(z) \log \frac{(1-z)^2}{z} + z - \frac{3}{2} \frac{(1-z)^2}{z} \right]. \end{aligned} \quad (51)$$





**Fig. 8:** Feynman diagrams giving  $gg$  real contributions in the infinite top-quark mass limit.

### 5.4.3 $gg \rightarrow Hg$

The calculation of the  $d$ -dimensional  $gg \rightarrow Hg$  amplitude involves the four diagrams shown in Fig. 8 and it is not so trivial to do by hand, yet the final result is very compact:

$$|\overline{\mathcal{M}}|^2 = \frac{1}{24(1-\epsilon)^2} \frac{\alpha_S^3}{\pi v^2} \frac{(m_H^8 + s^4 + t^4 + u^4)(1-2\epsilon) + \frac{1}{2}\epsilon(m_H^4 + s^2 + t^2 + u^2)^2}{stu}. \quad (52)$$

This example is illustrative of the fact that keeping track of the  $\epsilon$  parts in the amplitude squared makes the calculation significantly more complex for at least two reasons. First the structure of the result itself is more involved. Second, one is forced to work at the squared amplitude level as  $d$  dimensional contributions come from the  $(d-2)$  dimensional gluon polarizations and therefore cannot exploit the beauty, power and simplicity of helicity amplitude techniques [20, 21]. Computing QCD amplitudes where states have fixed polarizations entails huge simplifications and allows to make predictions for amplitudes with many external partons. For example, tree-level amplitudes in the HEFT involving up to 5 extra partons can be easily obtained automatically using tools such as ALPGEN [22] or MADGRAPH [23]. Fortunately, it turns out that is possible to use a different scheme than CDR and actually perform the computation of the Born and real matrix elements in exactly four dimensions (yet integrate them over the  $d$ -dimensional phase space). This involves a different (and a bit tricky)  $d$ -dimensional algebra for the loop computations and the introduction of (universal) finite terms for the initial-state counter-terms and UV subtractions, yet with an enormous computational simplification. All public NLO codes for processes at the LHC in practice do use such "maximally four dimensional"  $d$ -dimensional regularisation schemes. Integrating the amplitude (52) over the  $d$ -dimensional phase space of Eq. (48) gives

$$\begin{aligned} \hat{\sigma}_R(gg) = & \hat{\sigma}_0 \frac{\alpha_S}{2\pi} C_A \left( \frac{\mu^2}{m_H^2} \right)^\epsilon c_\Gamma \left[ \left( \frac{2}{\epsilon^2} + \frac{2}{\epsilon} \frac{b_0}{C_A} - \frac{\pi^2}{3} \right) \delta(1-z) \right. \\ & - \frac{2}{\epsilon} p_{gg}(z) - \frac{11}{3} \frac{(1-z)^3}{z} - 4 \frac{(1-z)^2(1+z^2) + z^2}{z(1-z)} \log z \\ & \left. + 4 \frac{1+z^4 + (1-z)^4}{z} \left( \frac{\log(1-z)}{1-z} \right)_+ \right], \quad (53) \end{aligned}$$

where the plus prescription is defined as follows:

$$\int_0^1 dx [h(x)]_+ f(x) = \int_0^1 dx h(x) [f(x) - f(1)]. \quad (54)$$

Note that the  $\frac{2}{\epsilon} \frac{b_0}{C_A} \delta(1-z)$  in Eq. (53) comes from reexpressing the divergent term  $-\frac{4}{\epsilon} [\frac{z}{(1-z)_+} + \frac{1-z}{z} + z(1-z)]$  in terms of  $-\frac{2}{\epsilon} p_{gg}(z)$ , see Eq. (67). The factorisation of the collinear divergence is handled by adding the corresponding counterterm

$$\hat{\sigma}_{\text{c.t.}}^{\text{coll.}}(gg) = 2 \hat{\sigma}_0 \frac{\alpha_S}{2\pi} \left[ \left( \frac{\mu^2}{\mu_F^2} \right)^\epsilon \frac{c_\Gamma}{\epsilon} P_{gg}(z) \right], \quad (55)$$

which gives

$$\begin{aligned} \hat{\sigma}_R^{\overline{\text{MS}}}(gg) &= \hat{\sigma}_R(gg) + \hat{\sigma}_{\text{c.t.}}^{\text{coll.}}(gg) \\ &= \hat{\sigma}_0 \frac{\alpha_S}{2\pi} C_A \left( \frac{\mu^2}{m_H^2} \right)^\epsilon c_\Gamma \left[ \left( \frac{2}{\epsilon^2} + \frac{2}{\epsilon} \frac{b_0}{C_A} - \frac{\pi^2}{3} \right) \delta(1-z) \right. \\ &\quad + 2p_{gg} \log \frac{m_H^2}{\mu_F^2} - \frac{11}{3} \frac{(1-z)^3}{z} - 4 \frac{(1-z)^2(1+z^2) + z^2}{z(1-z)} \log z \\ &\quad \left. + 4 \frac{1+z^4 + (1-z)^4}{z} \left( \frac{\log(1-z)}{1-z} \right)_+ \right]. \end{aligned} \quad (56)$$

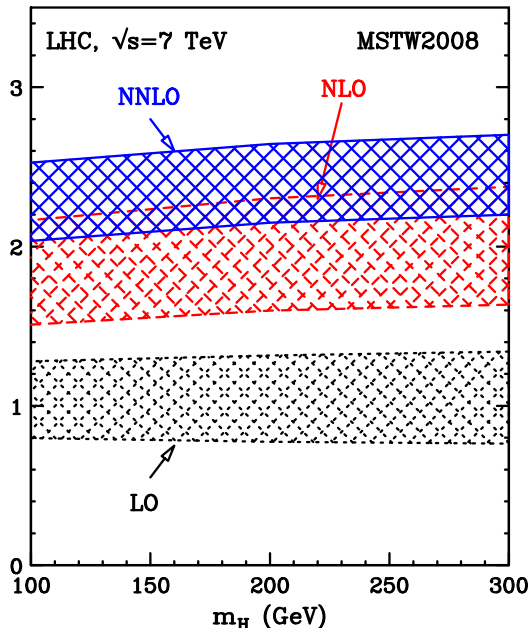
We can now recognise that the IR poles match those of the virtual contributions in Eq. (41). Adding up the contributions from real and virtual contributions of the  $gg$  channel we obtain (note that our definition of  $\sigma_0$ , Eq. (34), contains a factor  $z$ ):

$$\begin{aligned} \hat{\sigma}^{\overline{\text{MS}}}(gg) &= \hat{\sigma}_R^{\overline{\text{MS}}}(gg) + \hat{\sigma}_V^{\overline{\text{MS}}}(gg) \\ &= \sigma_0 \frac{\alpha_S}{2\pi} C_A \left[ \left( \frac{11}{3} + \frac{2}{3} \pi^2 - 2 \frac{b_0}{C_A} \log \frac{m_H^2}{\mu_R^2} \right) \delta(1-z) \right. \\ &\quad - \frac{11}{3} \frac{(1-z)^3}{z} + 2p_{gg} \log \frac{m_H^2}{\mu_F^2} - 4 \frac{(1-z+z^2)^2}{z(1-z)} \log z \\ &\quad \left. + 8 \frac{(1-z+z^2)^2}{z} \left( \frac{\log(1-z)}{1-z} \right)_+ \right]. \end{aligned} \quad (57)$$

As predicted, the final results for the short distance coefficients is finite (yet scheme dependent) and does contain the necessary  $\log$ 's of the renormalisation and factorisation scales that compensate up to  $\alpha_S^3$  the corresponding dependences in  $\alpha_S^2(\mu_R)$  of the Born amplitude and in the PDF's.

## 5.5 NLO results: discussion

The expressions above can be easily implemented in a numerical code to perform the convolution integrals with PDF's. A few simple numerical optimizations, such as the choice of integration variables, and a bit of attention to the implementation of the  $+$  distributions, that's all is needed. The reader can find a sample implementation in a Mathematica® notebook at the web address mentioned at the end of the Introduction. By running the code with different scale choices, one can associate an uncertainty to the NLO predictions as done at LO. The result, shown in Fig. 9, comes as a big surprise! The NLO calculation predicts a rate twice as large and the respective LO and NLO uncertainty bands do not even overlap. That means that our naive estimate of the uncertainties at LO is totally off and therefore unreliable. It seems also to suggest that perturbation expansion is at stake here. As we had mentioned, this motivated the computation of the NNLO corrections, which are also shown in Fig. 9. Fortunately, NNLO



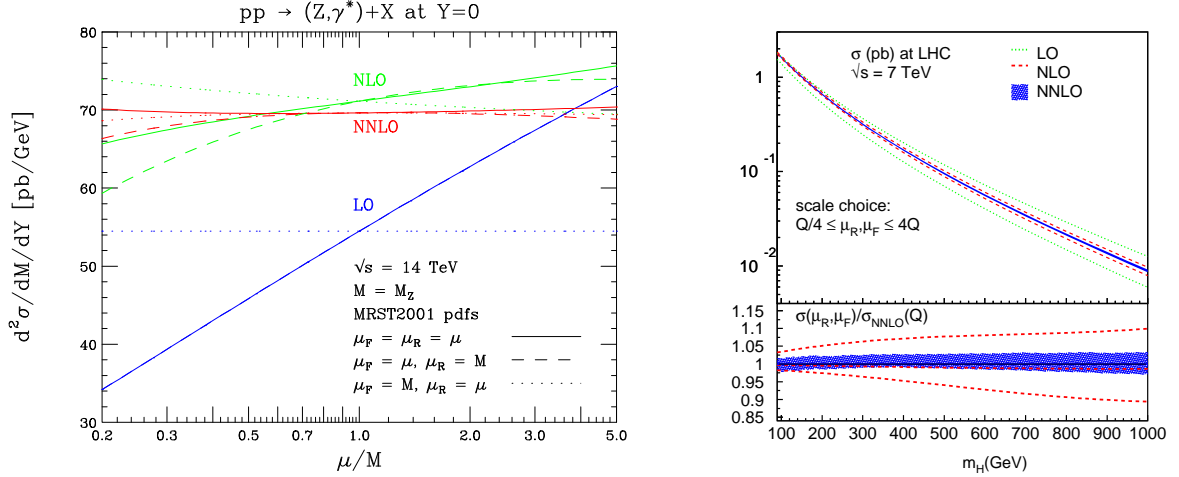
**Fig. 9:** K-factors for Higgs production from gluon fusion at the LHC. Uncertainty bands are obtained via independent scale variation  $1/2 m_H < \mu_R, \mu_F < 2 m_H$  with  $1/2 < \mu_F/\mu_R < 2$ . The LO and NLO bands can be obtained by implementing the formulas obtained in these notes in a code that performs the numerical integration over the PDF's. Cross-checks and NNLO results can be obtained with HNNLO [24]. (Plot courtesy of M. Grazzini).

predictions do overlap with NLO and also display a smaller scale dependence, so that the perturbation picture seems safe starting from NLO on. In fact, this behavior is rather special to  $pp \rightarrow H + X$  and it is often rephrased by saying that what we call LO (in the perturbative expansion) is not actually the leading one in size and therefore we should not start from that. For instance, in Drell-Yan or VBF this does not happen, and the perturbative expansions (seem to) converge beautifully, see Fig. 10. In any case, the Higgs production reminds us an important fact that we should always keep in mind: scale variation cannot by definition reproduce missing finite terms in the perturbative expansion and as such can only give an indication of what the real uncertainties could be. On the other hand, comparison between predictions from LO and NNLO, their stabilization (or lack thereof) and the use of approximate methods to determine (classes of) higher order terms, all together can provide a rather solid picture on the theoretical uncertainties on a case-by-case basis. We mention, in passing, another important source of uncertainties in making predictions for hadron colliders, i.e., that coming from imperfect knowledge of the PDF's. Uncertainties are related to unknown higher-order terms in the DGLAP evolution equations that determine as well as from the extraction of the initial condition from experimental data, see [QCD:3] and in particular [QCD:3.3.2].<sup>7</sup>

As far as total cross sections are concerned, the situation is therefore pretty clear. Fixed-order calculations come equipped with self-detecting procedures that can give us information on whether a prediction is reliable or not. If not, it can be systematically improved by including higher-order terms (almost for free nowadays at NLO, yet at a rather high cost at NNLO) and uncertainties can be easily estimated. So it is natural to ask, what about other IR-safe observables?

Let us consider, once again  $pp \rightarrow H + X$  as an example, and focus on the Higgs momentum

<sup>7</sup>The latter does in fact imply also the prediction of experimental observables at the same order in perturbation theory and therefore are also intrinsically also affected by scale dependencies. Such effects are not included normally in the estimation of the uncertainties coming from PDF's.

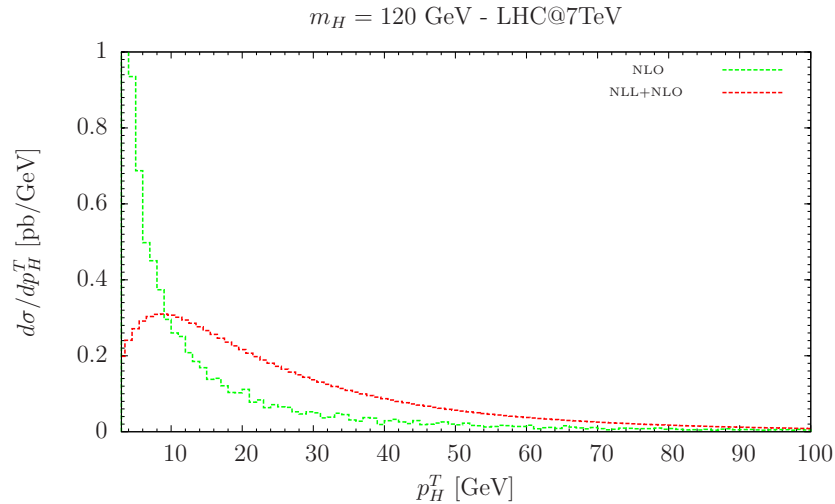


**Fig. 10:** Examples of improvement in the predictions of processes at LHC in going from LO to NNLO. On the left, scale dependence of the predictions for  $Z/\gamma^*$  production (at  $y = 0$ ) at the LHC14, at fixed order [25]. On the right, Higgs production at the LHC7 via VBF [26] as a function of the Higgs mass. The bands are obtained by independent scale variation in the interval  $Q/4 \leq \mu_F, \mu_R \leq 4Q$ ,  $Q$  being the virtuality of the  $W, Z$  fusing into the Higgs. In both cases the perturbative expansion behaves extremely well and NNLO predictions overlap with those at LO and NLO and display a much smaller residual uncertainty.

(fully inclusive) distribution, which can be parametrized in terms of only two variables<sup>8</sup>, the rapidity  $y_H$  and the transverse momentum  $p_H^T$ . At LO (referred to the total cross section), the Higgs can be boosted in the forward or backward directions in the lab system,  $y_H = \frac{1}{2} \log \frac{x_1}{x_2}$ , yet it has always  $p_H^T = 0$ , i.e. the distribution in  $p_H^T$  is a delta function centered at  $p_H^T = 0$ . At NLO (again referred to the total cross section),  $2 \rightarrow 2$  diagrams enter in the calculation and the Higgs can have a non-zero  $p_H^T$ . Since at any point in phase space with  $p_H^T \neq 0$  this is the first non-zero contribution, the observable  $p_H^T$  of the Higgs is only at LO. In other words if we want to know the  $p_H^T$  distribution of the Higgs at NLO over all phase space, we need at least a NNLO prediction for the cross section. Another way of thinking about it is to ask oneself what kind of diagrams are present in the calculation for that observable in a given area of the phase space: if there are only tree-level diagrams then the observable is LO. It is important when working with NLO codes to always think about what kind of observables are actually predicted at NLO, what at LO and what not even at LO. Again, a NNLO computation for the total cross section for  $pp \rightarrow H + X$ , gives NNLO information on the Higgs rapidity distribution, NLO for the Higgs  $p_H^T$  and  $pp \rightarrow H + 1$ -jet observables, LO for  $pp \rightarrow H + 2$ -jets observables and the structure of the jet in  $H + 1$ -jet events and no information at all on  $pp \rightarrow H + 3$ -jets observables. In short, a fixed-order computation can only make predictions for a finite number of observables, typically with a rather limited number of resolved partons and a very small number of unresolved ones, i.e. just one for a NLO computation and up to two for a NNLO computation. This is the first main limitation of a fixed-order computation. However, it is not the only one.

Consider again the  $p_H^T$  distribution of the Higgs as predicted by a NLO computation for the total cross section, Fig. 11. This curve can be easily obtained using the expressions in four dimensions of Eqs. (42,47,52), performing the integration over the polar angle together with the PDF's via a Monte-Carlo method and plotting it point-by-point during the integration. The  $p_H^T$  distribution is divergent in  $p_H^T = 0$  as expected from soft and beam-collinear emissions. As we have learnt such divergences are proportional to  $\delta(1-z)$  where  $z$  is the fraction of parton-parton energy taken by the Higgs and are cancelled by the virtual contributions, all of which reside in  $p_T = 0$ . So the cancellation between real and virtual contributions, all of it happens in the first bin of the histogram. How do we interpret such

<sup>8</sup>We do not consider the azimuthal angle  $\phi$ , because for symmetry reasons can only lead to a uniform distribution



**Fig. 11:** Higgs  $p_H^T$  spectrum for a Higgs of  $m_H = 120$  GeV at the LHC7. The labeling NLO and NLL+NLO refer to the total cross section. The curves are normalized to the same value (=total cross section is the same). The green curve is just a LO prediction for the  $p_H^T$  of the Higgs. The logarithmic divergence at  $p_H^T \rightarrow 0$  is cancelled by the negative infinite virtual contributions at  $p_H^T = 0$  (not shown!). The resummed prediction (red curve) features a “physical” smooth behavior at small  $p_H^T$ . (The resummed prediction is obtained via HqT [27]).

weird distribution? A useful way is to think about the size of the bin of the distribution as our resolution scale: with a rather coarse binning there is no “going-to-infinity” and predictions are rather stable (this of course includes the total cross section which corresponds to using only one bin), while with thin binning, we start to be sensitive to low energy and virtual emissions which become increasingly important and are not included at all in a fixed-order approach. This is the case where resummed predictions come into rescue: one finds that the leading part of soft emissions (real and virtual) is universal, it can be considered at all orders and included by identifying the log’s associated to it and exponentiating them. This can be done either at very high accuracy analytically yet fully inclusively or in a numerical and exclusive way at the leading log with a parton shower (which actually resums both soft and collinear enhancements). The result of including these effects analytically is shown in Fig. 11, red curve. In very crude words, the effect of the resummation is to spread the  $\delta(p_T)$  of the virtual contributions over a range of a few tens of GeV with the effect of smoothing out the divergence and producing a “physical” distribution.

In summary, fixed-order calculations in perturbative QCD can be performed in a well-defined and quite simple framework, i.e. in the context of the factorization theorem. It is therefore possible to make predictions for inclusive quantities in hadron colliders, which can be systematically improved at the “only” price of an (exponential) increase in the complexity of the calculation. In practice, however, the use of fixed-order predictions is limited by several other important drawbacks. First, only processes with a few resolved partons can be calculated, while in practice we know that hundreds of hadrons can be produced in a single proton-proton interaction of which we are bound to ignore the details. Second, sharp infinities appear in the phase that do cancel between real and virtual contributions if inclusive enough observables are defined, yet lead to unphysical distributions in specific areas of the phase space and/or when the resolved partons become either soft or collinear. Such local positive and negative infinities are unphysical because they appear only due the artificial truncation of the perturbative expansion. Finally, the fact that plus and minus infinities appear locally in phase space also means that fixed order predictions beyond LO cannot be used as probability functions to generate events as distributed in nature. Parton showers, i.e. fully exclusive resummation, and their merging/matching with fixed-order predictions, provide an elegant and powerful way out to all the above limitations.

## 6 Beyond fixed-order predictions

As we have explicitly verified, fixed-order predictions have important limitations both of principle (areas of phase space and observables, such as jet substructure are poorly described, no hadrons but only partons) and in practice (no event simulation is possible). Fortunately, an alternative approach exists that is based on the fact that the IR structure, soft or collinear, of QCD is universal and contributions can be resummed at all orders. Last but not least, formulas that describe the emission of soft and collinear partons are amenable of a probabilistic interpretation and therefore not only it is possible to perform an explicit resummation but also to associate a full “history” to an hard scattering event, i.e., to associate to every event a full-fledged description of an high-energy event from the two initial protons to the final (possibly hundreds) of hadrons and leptons in the final state. In addition, in the latest years, enormous progress has been achieved in combining the accuracy of fixed-order predictions with the flexibility of parton showers. These methods are briefly presented here together with their applications to Higgs production. The short presentation below is adapted from Ref. [28]. The reader is also referred to [QCD:4.4] for further details, examples and references.

### 6.1 Parton Showers

Parton Showers (PS) are able to dress a given Born process with all the dominant (i.e. enhanced by collinear logarithms, and to some extent also soft ones) QCD radiation processes at all orders in perturbation theory. In particular, the dominant contributions, i.e. those given by the leading logarithms, coming from both real and virtual emissions are included. The cross section for the first (which is often also the hardest) emission in a shower reads:

$$d\sigma^{\text{1st step}} = d\Phi_B B(\Phi_B) \left[ \Delta(p_{\perp}^{\text{min}}) + d\Phi_{R|B} \Delta(p_T(\Phi_{R|B})) P(\Phi_{R|B}) \right], \quad (58)$$

where  $\Delta(p_T)$  denotes the Sudakov form factor

$$\Delta(p_T) = \exp \left[ - \int d\Phi_{R|B} P(\Phi_{R|B}) \Theta(p_T(\Phi_R) - p_T) \right]. \quad (59)$$

This Sudakov form factor can be understood as a no-emission probability of secondary partons down to a resolution scale of  $p_T$ . Here  $P(\Phi_{R|B})$  is a process-independent universal splitting function that allows to write the PS approximation to the real cross section  $R^{\text{PS}}$ , typically given schematically by a product of the underlying Born-level term folded with a splitting kernel  $P$

$$R^{\text{PS}}(\Phi) = P(\Phi_{R|B}) B(\Phi_B). \quad (60)$$

In this framework,  $\Phi_{R|B}$  is often expressed in terms of three showering variables, like the virtuality  $t$  in the splitting process, the energy fraction of the splitting  $z$  and the azimuth  $\phi$ . A very simple (and widely used) choice for the splitting function, is

$$P(\Phi_{R|B}) d\Phi_{R|B} = \frac{\alpha_S(t)}{2\pi} P_{a \rightarrow bc}(z) \frac{d\phi}{2\pi} \frac{dt}{t} dz \quad (61)$$

where  $P(z)$  are Altarelli-Parisi splitting functions on which any QCD amplitude factorises in the collinear limit  $b \parallel c$ .

The above definition of the Sudakov form factor, guarantees that the square bracket in Eq. (58) integrates to unity, a manifestation of the probabilistic nature of the parton shower. Thus, integrating the shower cross section over the radiation variables yields the total cross section, given at LO by the Born amplitude. The corresponding radiation pattern consists of two parts: one given by the first term in the square bracket, where no further resolvable emission above the parton-shower cut-off  $p_{\perp}^{\text{min}}$  – typically of the order of 1 GeV – emerges, and the other given by the second term in the square bracket describing

the first emission, as determined by the splitting kernel. It is important to stress that the real-emission cross section in a PS generator is only correct in the small angle and/or soft limit, where  $R^{\text{PS}}$  is a reliable approximation of the complete matrix element.

After the 1st step the process is repeated using the new configuration as the Born one.

While rather crude, the PS approximation is a very powerful one, due mainly to the great flexibility and simplicity in the implementation of  $2 \rightarrow 1$  and  $2 \rightarrow 2$  high- $Q^2$  processes. In addition, once augmented with a hadronisation model the simulation can easily provide a full description of a collision in terms of physical final states, i.e., hadrons, leptons and photons. In the current terminology a generic Monte Carlo generator mainly refers to such tools, the most relevant examples of are PYTHIA 6 and PYTHIA 8 [29, 30], HERWIG [31], HERWIG++ [32], and SHERPA [33]. A very clear and exhaustive presentation of parton shower generators can be found in Ref. [34].

## 6.2 Matrix-element merging (ME+PS)

In parton showers algorithms QCD radiation is generated in the collinear and soft approximation, using Markov chain techniques based on Sudakov form factors. Hard and widely separated jets are thus poorly described in this approach. On the other hand, tree-level fixed order amplitudes can provide reliable predictions in the hard region, while failing in the collinear and soft limits. To combine both descriptions and avoid double counting or gaps between samples with different multiplicity, an appropriate merging method is required.

Matrix-element merging [35] aims at correcting as many large-angle emissions as possible with the corresponding tree-level accurate prediction, rather than only *small-angle* accurate. This is achieved by generating events up to a given (high) multiplicity using a matrix-element generator, with some internal jet-resolution parameter  $Q_{\text{cut}}$  on the jet separation, such that practically all emissions above this scale are described by corresponding tree-level matrix elements. Their contributions are corrected for running-coupling effects and by Sudakov form factors. Radiation below  $Q_{\text{cut}}$  on the other hand is generated by a parton-shower program, which is required to veto radiation with separation larger than  $Q_{\text{cut}}$ . As far as the hardest emission is concerned, matrix-element merging is as accurate as matrix-element corrections (when these are available) or NLO+PS. Since they lack NLO virtual corrections, however, they do not reach NLO accuracy for inclusive quantities. Nevertheless, they are capable to achieve leading-order accuracy for multiple hard radiation, beyond the hardest only, while NLO+PS programs, relying on the parton shower there are only accurate in the collinear and/or soft limit for these quantities.

Several merging schemes have been proposed, which include the CKKW scheme [35–37] and its improvements [38, 39], the MLM matching [40], and the  $k_T$ -MLM variation [41]. The MLM schemes have been implemented in several matrix element codes such as ALPGEN [22], MADGRAPH [23], through interfaces to PYTHIA/HERWIG, while SHERPA [33] and HERWIG++ [32] have adopted the CKKW schemes and rely on their own parton showers. In Ref. [42] a detailed, although somewhat outdated description of each method has been given and a comparative study has been performed.

## 6.3 NLO+PS in a nutshell

Several proposals have been made for the full inclusion of complete NLO effects in PS generators. At this moment, only two of them have reached a mature enough stage to be used in practice: MC@NLO [43] and POWHEG [44]. Both methods correct – in different ways – the real-emission matrix element to achieve an exact tree-level emission matrix element, even at large angle. As we have seen in the previous subsection, this is what is also achieved with matrix-element corrections in parton showers, at least for the simplest processes listed earlier. This, however, is not sufficient for the NLO accuracy, since the effect of virtual corrections also needs to be included. In both methods, the real-emission cross section is split into a singular and non-singular part,  $R = R^s + R^f$ . One then computes the total NLO inclusive

cross section, excluding the finite contribution, at fixed underlying Born kinematics, defined as

$$\bar{B}^s = B(\Phi_B) + \left[ V(\Phi_B) + \int d\Phi_{R|B} R^s(\Phi_{R|B}) \right], \quad (62)$$

and uses the formula

$$d\sigma^{\text{NLO+PS}} = d\Phi_B \bar{B}^s(\Phi_B) \left[ \Delta^s(p_{\perp}^{\min}) + d\Phi_{R|B} \frac{R^s(\Phi_R)}{B(\Phi_B)} \Delta^s(p_T(\Phi)) \right] + d\Phi_R R^f(\Phi_R) \quad (63)$$

for the generation of the events. In this formula, the term  $\bar{B}$  can be understood as a local  $K$ -factor reweighting the soft matrix-element correction part of the simulation. Clearly, employing the fact that the term in the first square bracket integrates to unity, as before, the cross section integrates to the full NLO cross section.

In MC@NLO one chooses  $R^s$  to be identically equal to the term  $B \otimes P$  that the PS generator employs to generate emissions. Within MC@NLO,  $n$ -body events are obtained using the  $\bar{B}^s$  function, and then fed to the PS, which will generate the hardest emission according to Eq. (62). These are called  $\mathcal{S}$  events in the MC@NLO language. An appropriate number of events are also generated according to the  $R^f$  cross section, and are directly passed to the PS generator. These are called  $\mathcal{H}$  events. In MC@NLO,  $R^f = R - R^s$  is not positive definite, and it is thus necessary to generate negative weighted events in this framework. A library of MC@NLO Higgs processes (gluon fusion, vector-boson associated production, and charged Higgs associated with top) is available at Ref. [45], which is interfaced to HERWIG and HERWIG++. A fully automatized approach, AMC@NLO [46] implemented in the MADGRAPH framework, is now available that allows to compute and combine all necessary ingredients (Born, real, virtual matrix elements plus counterterms) at the user's request.

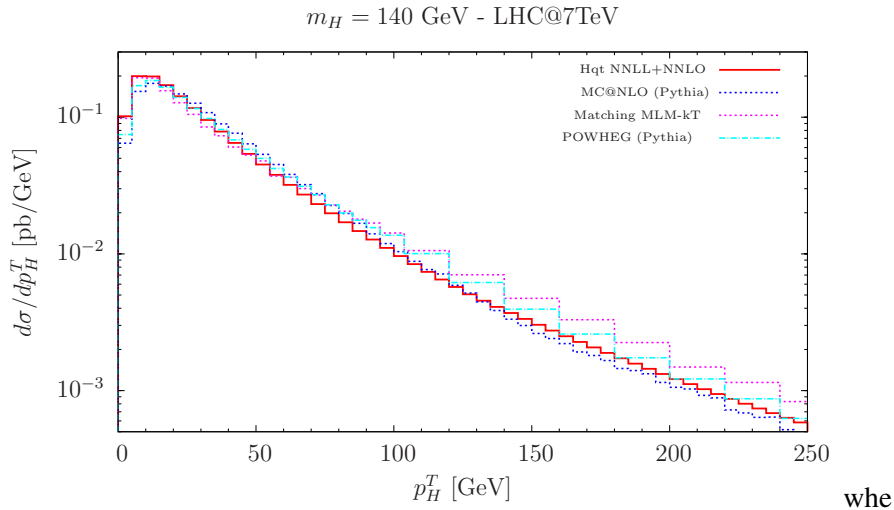
In POWHEG, one chooses  $R^s \leq R$ , and in many cases even  $R^s = R$ , so that the finite cross section  $R^f$  vanishes. In this case, the hardest emission is generated within POWHEG itself, and the process is passed to the parton shower only after the hardest radiation is generated. Positive weighted events are obtained, since  $R^f$  can always be chosen to be positive definite. In all cases the chosen  $R^s$  has exactly the same singularity structure as  $R$ , so that  $R^f$  always yield a finite contribution to the cross section. Implementations of Higgs production processes with the POWHEG method are available in HERWIG++ [47], in the POWHEGBOX [48] (interfaced to both HERWIG and PYTHIA) and recently in SHERPA [49].

## 6.4 Improved descriptions of Higgs production

Being of primary importance, Higgs kinematic distributions are now quite well predicted and also available via public codes such as ResBos [50] and HqT [27, 51]. Differential  $p_H^T$  distributions accurate to LO yet featuring the exact bottom- and top-quarks mass loop dependence (and therefore can be used also for predictions of scalar Higgs in BSM) can be obtained via HIGLU [52] as well as via HPro [53]. However, in experimental analyses, it is also crucial to get as precise predictions as possible for exclusive observables that involve extra jets, such as the jet  $p_T$  spectra and the jet rates, at both parton and hadron level. To optimize the search strategies and in particular to curb the very large backgrounds, current analyses both at Tevatron and at the LHC select 0-, 1- and 2-jet events and perform independent analyses on each sample. The final systematic uncertainties are effected by both the theoretical and experimental ones of such a jet-bin based separation. In the HEFT, fully exclusive parton- and hadron-level calculations can now be performed by Parton Shower (PS) programs or with NLO QCD codes matched with parton showers: via the MC@NLO and POWHEG methods. Beyond the HEFT, fully exclusive predictions ME+PS and NLO+PS techniques has become available only recently [54, 55]. The reason is that one needs to compromise between the validity of HEFT and the complexity of higher loop calculations.

Fig. 12 shows a comparison of the predictions of the  $p^T$  of the Higgs at LHC7 as obtained in HEFT from:





**Fig. 12:** Higgs  $p_H^T$  spectrum for a Higgs of  $m_H = 140$  GeV as predicted by a series of improved predictions: NNLL+NNLO resummed (red solid), MC@NLO + Pythia (blue dashes), matrix-element + Pythia merged results (magenta dashes), POWHEG + Pythia (cyan dashes). All predictions display similar features, i.e. a peak between 10-20 GeV and a similar shape at high- $p_H^T$  with differences that lie within their respective uncertainties (not shown).

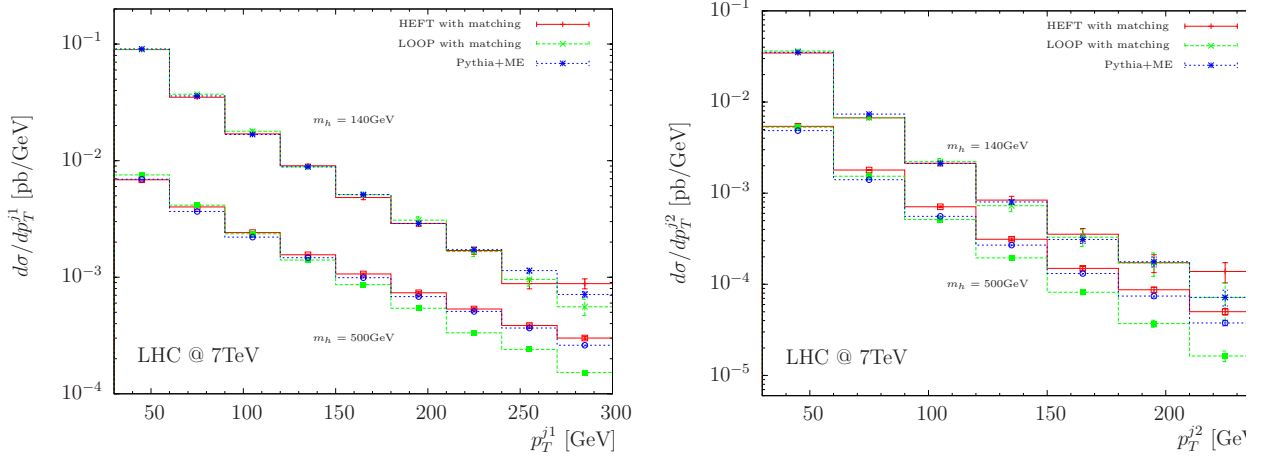
- a full analytical resummation at NNLL;
- MC@NLO (w/ PYTHIA);
- ME+PS merging (MADGRAPH+PYTHIA);
- POWHEG (w/ PYTHIA).

We first stress (again) that this observable which is at NLO at high- $p^T$  only in the Hqt predictions. The ME+PS approach is built to be LO for all observables, while MC@NLO and POWHEG predictions are based on the NLO calculation for the total cross section, the same performed in these notes. Notwithstanding we see that given the expected uncertainties which are quite large above all at high- $p^T$  the shapes are in substantial agreement both in the low and high- $p^T$  ranges. In Fig. 13 the  $p^T$  distributions for the first and second jets are shown comparing the ME+PS prediction based on the HEFT and one with the full top-mass dependence and PYTHIA. Even in this case the agreement between the various approaches is extremely good for a light Higgs. For a very heavy Higgs difference in the  $p_T$  distributions of the extra jets become visible at quite a high  $p^T$ , a region not very relevant phenomenologically.

## 7 Conclusions

Progress in the field of QCD predictions for the LHC in the form of MC tools usable by both theorists and experimentalists has made tremendous progress in the last years. It is fair to say that we are now able (or close to be able in some specific very challenging cases) to compute automatically or semi-automatically any interesting cross section for Standard Model and Beyond processes at NLO accuracy and interface it with parton shower programs for event generation. In the LHC era the lowest acceptable accuracy for any serious phenomenological and experimental study is via an NLO event generator. LHC precision physics is now at NNLO in QCD and NLO in EW. Any physicist interested in making discoveries at the LHC needs to be familiar with the ideas, the physics and the reach of the current QCD simulation tools.

To this aim, we have considered  $pp \rightarrow H + X$  as a case study. We have illustrated how accurate and useful predictions for cross sections and other observables can be obtained in QCD, starting from the calculation of Born amplitude (at one loop) and the corresponding hadronic cross section. We have



**Fig. 13:** Jet  $p_T$  distributions for associated jets in gluon fusion production of  $m_H = 140$  GeV and  $m_H = 500$  GeV Higgs bosons at 7 TeV LHC.

then considered Higgs production at NLO in the HEFT and discussed the limitations of fixed-order predictions. Finally, we have briefly discussed how fully exclusive predictions are obtained with modern tools, that allow to reach the accuracy of NLO predictions together with the full exclusivity of a parton shower approach.

## Appendix

### Splitting functions and collinear counterterms

We define the 4-dimensional splitting functions as in (4.94) of the ESW book:

$$P_{qq}(z) = C_F p_{qq}(z) = C_F \left[ \frac{1+z^2}{(1-z)_+} + \frac{3}{2} \delta(1-z) \right] \quad (64)$$

$$P_{qg}(z) = T_R p_{qg}(z) = T_R [z^2 + (1-z)^2] \quad (65)$$

$$P_{gq}(z) = C_F p_{gq}(z) = C_F \left[ \frac{1+(1-z)^2}{z} \right] \quad (66)$$

$$P_{gg}(z) = C_A p_{gg}(z) = 2C_A \left[ \frac{z}{(1-z)_+} + \frac{1-z}{z} + z(1-z) \right] + b_0 \delta(1-z), \quad (67)$$

where  $b_0 = 11/6 C_A - 2n_f T_F/3$ . We also define the following quantities as the extension of the splitting functions in  $d$ -dimensions:

$$P_{ij}^d(z) = P_{ij}(z) + \epsilon P_{ij}^\epsilon(z) \quad (68)$$

where

$$P_{qq}^\epsilon(z) = C_F p_{qq}^\epsilon(z) = -C_F(1-z) \quad (69)$$

$$P_{qg}^\epsilon(z) = T_R p_{qg}^\epsilon(z) = -T_R 2z(1-z) \quad (70)$$

$$P_{gq}^\epsilon(z) = C_F p_{gq}^\epsilon(z) = -C_F z \quad (71)$$

$$P_{gg}^\epsilon(z) = 0 \quad (72)$$

factorisation of the collinear divergences is performed through the addition of the following counterterm for each parton in the initial state:

$$\sigma_{\text{c.t.}}^{\text{CDR}} = \sigma_0^{\text{CDR}} \frac{\alpha_S}{2\pi} \left[ \left( \frac{\mu^2}{\mu_F^2} \right)^\epsilon \frac{c_\Gamma}{\epsilon} P_{ij}(z) \right] \quad (73)$$

where  $\sigma_0^{\text{SCHEME}}$  is the LO cross section and its value depends on the scheme (see the example for Drell-Yan)]. In CDR, when there is a collinear divergence, the cross section behaves as

$$\sigma_R^{\text{coll}} \sim -\frac{1}{\epsilon} P_{ij}^d(z) \sigma_0^{\text{CDR}} + \text{other terms.} \quad (74)$$

Adding the counterterm (73), leaves a finite part

$$\sigma_R^{\overline{\text{MS}}} \sim -P_{ij}^e(z) (\sigma_0^{\text{CDR}}|_{\epsilon \rightarrow 0}) + \text{other terms.} \quad (75)$$

## References

- [1] R. K. Ellis, W. J. Stirling, and B. Webber, “QCD and collider physics,” *Camb.Monogr.Part.Phys.Nucl.Phys.Cosmol.*, vol. 8, pp. 1–435, 1996.
- [2] M. L. Mangano, “Introduction to QCD,” <http://cdsweb.cern.ch/record/454171/files/open-2000-255.pdf>, no. CERN-OPEN-2000-255, 1999.
- [3] P. Nason, “Introduction to QCD,” <http://doc.cern.ch/cernrep/1998/98-03/98-03.html>, vol. C9705251, pp. 94–149, 1997.
- [4] G. P. Salam, “Perturbative QCD for the LHC,” *PoS*, vol. ICHEP2010, p. 556, 2010.
- [5] H. Georgi, S. Glashow, M. Machacek, and D. V. Nanopoulos, “Higgs Bosons from Two Gluon Annihilation in Proton Proton Collisions,” *Phys.Rev.Lett.*, vol. 40, p. 692, 1978.
- [6] S. Dawson, “Radiative corrections to Higgs boson production,” *Nucl.Phys.*, vol. B359, pp. 283–300, 1991.
- [7] A. Djouadi, M. Spira, and P. Zerwas, “Production of Higgs bosons in proton colliders: QCD corrections,” *Phys.Lett.*, vol. B264, pp. 440–446, 1991.
- [8] D. Graudenz, M. Spira, and P. Zerwas, “QCD corrections to Higgs boson production at proton proton colliders,” *Phys.Rev.Lett.*, vol. 70, pp. 1372–1375, 1993.
- [9] M. Spira, A. Djouadi, D. Graudenz, and P. Zerwas, “Higgs boson production at the LHC,” *Nucl.Phys.*, vol. B453, pp. 17–82, 1995.
- [10] R. V. Harlander and W. B. Kilgore, “Next-to-next-to-leading order Higgs production at hadron colliders,” *Phys.Rev.Lett.*, vol. 88, p. 201801, 2002.
- [11] C. Anastasiou and K. Melnikov, “Higgs boson production at hadron colliders in NNLO QCD,” *Nucl.Phys.*, vol. B646, pp. 220–256, 2002.
- [12] V. Ravindran, J. Smith, and W. L. van Neerven, “NNLO corrections to the total cross-section for Higgs boson production in hadron hadron collisions,” *Nucl.Phys.*, vol. B665, pp. 325–366, 2003.
- [13] R. V. Harlander, H. Mantler, S. Marzani, and K. J. Ozeren, “Higgs production in gluon fusion at next-to-next-to-leading order QCD for finite top mass,” *Eur. Phys. J.*, vol. C66, pp. 359–372, 2010.
- [14] A. Pak, M. Rogal, and M. Steinhauser, “Finite top quark mass effects in NNLO Higgs boson production at LHC,” *JHEP*, vol. 02, p. 025, 2010.
- [15] R. V. Harlander, F. Hofmann, and H. Mantler, “Supersymmetric Higgs production in gluon fusion,” *JHEP*, vol. 02, p. 055, 2011.
- [16] A. Pak, M. Rogal, and M. Steinhauser, “Production of scalar and pseudo-scalar Higgs bosons to next-to-next-to-leading order at hadron colliders,” *JHEP*, vol. 1109, p. 088, 2011.
- [17] M. R. Whalley, D. Bourilkov, and R. C. Group, “The Les Houches Accord PDFs (LHAPDF) and Lhaglu,” 2005.
- [18] S. Catani and M. Seymour, “A General algorithm for calculating jet cross-sections in NLO QCD,” *Nucl.Phys.*, vol. B485, pp. 291–419, 1997.
- [19] S. Frixione, Z. Kunszt, and A. Signer, “Three jet cross-sections to next-to-leading order,” *Nucl.Phys.*, vol. B467, pp. 399–442, 1996.

- [20] M. L. Mangano and S. J. Parke, “Multi-Parton Amplitudes in Gauge Theories,” *Phys. Rept.*, vol. 200, pp. 301–367, 1991.
- [21] L. J. Dixon, “Calculating scattering amplitudes efficiently,” 1996.
- [22] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, and A. D. Polosa, “ALPGEN, a generator for hard multiparton processes in hadronic collisions,” *JHEP*, vol. 0307, p. 001, 2003.
- [23] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, “MadGraph 5 : Going Beyond,” *JHEP*, vol. 1106, p. 128, 2011.
- [24] S. Catani and M. Grazzini, “An NNLO subtraction formalism in hadron collisions and its application to Higgs boson production at the LHC,” *Phys. Rev. Lett.*, vol. 98, p. 222002, 2007.
- [25] C. Anastasiou, L. J. Dixon, K. Melnikov, and F. Petriello, “High precision QCD at hadron colliders: Electroweak gauge boson rapidity distributions at NNLO,” *Phys.Rev.*, vol. D69, p. 094008, 2004.
- [26] P. Bolzoni, F. Maltoni, S.-O. Moch, and M. Zaro, “Vector boson fusion at NNLO in QCD: SM Higgs and beyond,” *Phys.Rev.*, vol. D85, p. 035002, 2012. 56 pages.
- [27] D. de Florian, G. Ferrera, M. Grazzini, and D. Tommasini, “Transverse-momentum resummation: Higgs boson production at the Tevatron and the LHC,” *JHEP*, vol. 11, p. 064, 2011.
- [28] S. Dittmaier *et al.*, “Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables,” 2011. Long author list - awaiting processing.
- [29] T. Sjostrand, S. Mrenna, and P. Z. Skands, “PYTHIA 6.4 Physics and Manual,” *JHEP*, vol. 0605, p. 026, 2006.
- [30] T. Sjostrand, S. Mrenna, and P. Z. Skands, “A Brief Introduction to PYTHIA 8.1,” *Comput.Phys.Commun.*, vol. 178, pp. 852–867, 2008.
- [31] G. Corcella, I. Knowles, G. Marchesini, S. Moretti, K. Odagiri, *et al.*, “HERWIG 6: An Event generator for hadron emission reactions with interfering gluons (including supersymmetric processes),” *JHEP*, vol. 0101, p. 010, 2001.
- [32] M. Bahr, S. Gieseke, M. Gigg, D. Grellscheid, K. Hamilton, *et al.*, “Herwig++ Physics and Manual,” *Eur.Phys.J.*, vol. C58, pp. 639–707, 2008. 143 pages, program and additional information available from <http://projects.hepforge.org/herwig>.
- [33] T. Gleisberg, S. Hoeche, F. Krauss, A. Schalicke, S. Schumann, *et al.*, “SHERPA 1. alpha: A Proof of concept version,” *JHEP*, vol. 0402, p. 056, 2004.
- [34] A. Buckley, J. Butterworth, S. Gieseke, D. Grellscheid, S. Hoche, *et al.*, “General-purpose event generators for LHC physics,” *Phys.Rept.*, vol. 504, pp. 145–233, 2011.
- [35] S. Catani, F. Krauss, R. Kuhn, and B. Webber, “QCD matrix elements + parton showers,” *JHEP*, vol. 0111, p. 063, 2001.
- [36] F. Krauss, “Matrix elements and parton showers in hadronic interactions,” *JHEP*, vol. 0208, p. 015, 2002.
- [37] L. Lonnblad, “Correcting the color dipole cascade model with fixed order matrix elements,” *JHEP*, vol. 0205, p. 046, 2002.
- [38] S. Hoeche, F. Krauss, S. Schumann, and F. Siegert, “QCD matrix elements and truncated showers,” *JHEP*, vol. 0905, p. 053, 2009.
- [39] K. Hamilton, P. Richardson, and J. Tully, “A Modified CKKW matrix element merging approach to angular-ordered parton showers,” *JHEP*, vol. 0911, p. 038, 2009.
- [40] M. L. Mangano, M. Moretti, and R. Pittau, “Multijet matrix elements and shower evolution in hadronic collisions:  $Wb\bar{b} + n$  jets as a case study,” *Nucl.Phys.*, vol. B632, pp. 343–362, 2002.
- [41] J. Alwall, S. de Visscher, and F. Maltoni, “QCD radiation in the production of heavy colored particles at the LHC,” *JHEP*, vol. 0902, p. 017, 2009.
- [42] J. Alwall, S. Hoche, F. Krauss, N. Lavesson, L. Lonnblad, *et al.*, “Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions,”

- Eur.Phys.J.*, vol. C53, pp. 473–500, 2008.
- [43] S. Frixione and B. R. Webber, “Matching NLO QCD computations and parton shower simulations,” *JHEP*, vol. 0206, p. 029, 2002.
- [44] P. Nason, “A New method for combining NLO QCD with shower Monte Carlo algorithms,” *JHEP*, vol. 0411, p. 040, 2004.
- [45] S. Frixione, F. Stoeckli, P. Torrielli, B. R. Webber, and C. D. White, “The MCanLO 4.0 Event Generator,” 2010.
- [46] R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, R. Pittau, *et al.*, “Scalar and pseudoscalar Higgs production in association with a top-antitop pair,” *Phys.Lett.*, vol. B701, pp. 427–433, 2011.
- [47] K. Hamilton, P. Richardson, and J. Tully, “A Positive-Weight Next-to-Leading Order Monte Carlo Simulation for Higgs Boson Production,” *JHEP*, vol. 0904, p. 116, 2009.
- [48] S. Alioli, P. Nason, C. Oleari, and E. Re, “A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX,” *JHEP*, vol. 1006, p. 043, 2010.
- [49] S. Hoche, F. Krauss, M. Schonherr, and F. Siegert, “Automating the POWHEG method in Sherpa,” *JHEP*, vol. 1104, p. 024, 2011.
- [50] C. Balazs, J. Huston, and I. Puljak, “Higgs production: A Comparison of parton showers and resummation,” *Phys. Rev.*, vol. D63, p. 014021, 2001.
- [51] G. Bozzi, S. Catani, D. de Florian, and M. Grazzini, “Transverse-momentum resummation and the spectrum of the Higgs boson at the LHC,” *Nucl. Phys.*, vol. B737, pp. 73–120, 2006.
- [52] U. Langenegger, M. Spira, A. Starodumov, and P. Trubeb, “SM and MSSM Higgs Boson Production: Spectra at large transverse Momentum,” *JHEP*, vol. 0606, p. 035, 2006.
- [53] C. Anastasiou, S. Bucherer, and Z. Kunszt, “HPro: A NLO Monte-Carlo for Higgs production via gluon fusion with finite heavy quark masses,” *JHEP*, vol. 0910, p. 068, 2009.
- [54] J. Alwall, Q. Li, and F. Maltoni, “Matched predictions for Higgs production via heavy-quark loops in the SM and beyond,” *Phys. Rev.*, vol. D85, p. 014031, 2012.
- [55] E. Bagnaschi, G. Degrossi, P. Slavich, and A. Vicini, “Higgs production via gluon fusion in the POWHEG approach in the SM and in the MSSM,” *JHEP*, vol. 02, p. 088, 2012.



## Flavour Physics and CP Violation

Y. Nir

Department of Particle Physics and Astrophysics  
Weizmann Institute of Science, Israel

### Abstract

We explain the many reasons for the interest in flavor physics. We describe flavor physics and the related CP violation within the Standard Model, and explain how the B-factories proved that the Kobayashi-Maskawa mechanism dominates the CP violation that is observed in meson decays. We explain the implications of flavor physics for new physics, with emphasis on the “new physics flavor puzzle”, and present the idea of minimal flavor violation as a possible solution. We explain why the values flavor parameters of the Standard Model are puzzling, present the Froggatt-Nielsen mechanism as a possible solution, and describe how measurements of neutrino parameters are interpreted in the context of this puzzle. We show that the recently discovered Higgs-like boson may provide new opportunities for making progress on the various flavor puzzles.

### 1 What is flavor?

The term “**flavors**” is used, in the jargon of particle physics, to describe several copies of the same gauge representation, namely several fields that are assigned the same quantum charges. Within the Standard Model, when thinking of its unbroken  $SU(3)_C \times U(1)_{EM}$  gauge group, there are four different types of particles, each coming in three flavors:

- Up-type quarks in the  $(3)_{+2/3}$  representation:  $u, c, t$ ;
- Down-type quarks in the  $(3)_{-1/3}$  representation:  $d, s, b$ ;
- Charged leptons in the  $(1)_{-1}$  representation:  $e, \mu, \tau$ ;
- Neutrinos in the  $(1)_0$  representation:  $\nu_1, \nu_2, \nu_3$ .

The term “**flavor physics**” refers to interactions that distinguish between flavors. By definition, gauge interactions, namely interactions that are related to unbroken symmetries and mediated therefore by massless gauge bosons, do not distinguish among the flavors and do not constitute part of flavor physics. Within the Standard Model, flavor-physics refers to the weak and Yukawa interactions.

The term “**flavor parameters**” refers to parameters that carry flavor indices. Within the Standard Model, these are the nine masses of the charged fermions and the four “mixing parameters” (three angles and one phase) that describe the interactions of the charged weak-force carriers ( $W^\pm$ ) with quark-antiquark pairs. If one augments the Standard Model with Majorana mass terms for the neutrinos, one should add to the list three neutrino masses and six mixing parameters (three angles and three phases) for the  $W^\pm$  interactions with lepton-antilepton pairs.

The term “**flavor universal**” refers to interactions with couplings (or to parameters) that are proportional to the unit matrix in flavor space. Thus, the strong and electromagnetic interactions are flavor-universal. An alternative term for “flavor-universal” is “**flavor-blind**”.

The term “**flavor diagonal**” refers to interactions with couplings (or to parameters) that are diagonal, but not necessarily universal, in the flavor space. Within the Standard Model, the Yukawa interactions of the Higgs particle are flavor diagonal.

The term “**flavor changing**” refers to processes where the initial and final flavor-numbers (that is, the number of particles of a certain flavor minus the number of anti-particles of the same flavor) are

different. In “flavor changing charged current” processes, both up-type and down-type flavors, and/or both charged lepton and neutrino flavors are involved. Examples are (i) muon decay via  $\mu \rightarrow e\bar{\nu}_i\nu_j$ , and (ii)  $K^- \rightarrow \mu^-\bar{\nu}_j$  (which corresponds, at the quark level, to  $s\bar{u} \rightarrow \mu^-\bar{\nu}_j$ ). Within the Standard Model, these processes are mediated by the  $W$ -bosons and occur at tree level. In “**flavor changing neutral current**” (FCNC) processes, either up-type or down-type flavors but not both, and/or either charged lepton or neutrino flavors but not both, are involved. Example are (i) muon decay via  $\mu \rightarrow e\gamma$  and (ii)  $K_L \rightarrow \mu^+\mu^-$  (which corresponds, at the quark level, to  $s\bar{d} \rightarrow \mu^+\mu^-$ ). Within the Standard Model, these processes do not occur at tree level, and are often highly suppressed.

Another useful term is “**flavor violation**”. We explain it later in these lectures.

## 2 Why is flavor physics interesting?

- Flavor physics can discover new physics or probe it before it is directly observed in experiments. Here are some examples from the past:
  - The smallness of  $\frac{\Gamma(K_L \rightarrow \mu^+\mu^-)}{\Gamma(K^+ \rightarrow \mu^+\nu)}$  led to predicting a fourth (the charm) quark;
  - The size of  $\Delta m_K$  led to a successful prediction of the charm mass;
  - The size of  $\Delta m_B$  led to a successful prediction of the top mass;
  - The measurement of  $\varepsilon_K$  led to predicting the third generation.
  - The measurement of neutrino flavor transitions led to the discovery of neutrino masses.
- CP violation is closely related to flavor physics. Within the Standard Model, there is a single CP violating parameter, the Kobayashi-Maskawa phase  $\delta_{KM}$  [1]. Baryogenesis tells us, however, that there must exist new sources of CP violation. Measurements of CP violation in flavor changing processes might provide evidence for such sources.
- The fine-tuning problem of the Higgs mass, and the puzzle of the dark matter imply that there exists new physics at, or below, the TeV scale. If such new physics had a generic flavor structure, it would contribute to flavor changing neutral current (FCNC) processes orders of magnitude above the observed rates. The question of why this does not happen constitutes the *new physics flavor puzzle*.
- Most of the charged fermion flavor parameters are small and hierarchical. The Standard Model does not provide any explanation of these features. This is the *Standard Model flavor puzzle*. The puzzle became even deeper after neutrino masses and mixings were measured because, so far, neither smallness nor hierarchy in these parameters have been established.

## 3 Flavor in the Standard Model

A model of elementary particles and their interactions is defined by the following ingredients: (i) The symmetries of the Lagrangian and the pattern of spontaneous symmetry breaking; (ii) The representations of fermions and scalars. The Standard Model (SM) is defined as follows:

(i) The gauge symmetry is

$$G_{\text{SM}} = SU(3)_C \times SU(2)_L \times U(1)_Y. \quad (1)$$

It is spontaneously broken by the VEV of a single Higgs scalar,  $\phi(1, 2)_{1/2}$  ( $\langle \phi^0 \rangle = v/\sqrt{2}$ ):

$$G_{\text{SM}} \rightarrow SU(3)_C \times U(1)_{\text{EM}}. \quad (2)$$

(ii) There are three fermion generations, each consisting of five representations of  $G_{\text{SM}}$ :

$$Q_{Li}(3, 2)_{+1/6}, \quad U_{Ri}(3, 1)_{+2/3}, \quad D_{Ri}(3, 1)_{-1/3}, \quad L_{Li}(1, 2)_{-1/2}, \quad E_{Ri}(1, 1)_{-1}. \quad (3)$$



### 3.1 The interaction basis

The Standard Model Lagrangian,  $\mathcal{L}_{\text{SM}}$ , is the most general renormalizable Lagrangian that is consistent with the gauge symmetry (1), the particle content (3) and the pattern of spontaneous symmetry breaking (2). It can be divided to three parts:

$$\mathcal{L}_{\text{SM}} = \mathcal{L}_{\text{kinetic}} + \mathcal{L}_{\text{Higgs}} + \mathcal{L}_{\text{Yukawa}}. \quad (4)$$

As concerns the kinetic terms, to maintain gauge invariance, one has to replace the derivative with a covariant derivative:

$$D^\mu = \partial^\mu + ig_s G_a^\mu L_a + ig W_b^\mu T_b + ig' B^\mu Y. \quad (5)$$

Here  $G_a^\mu$  are the eight gluon fields,  $W_b^\mu$  the three weak interaction bosons and  $B^\mu$  the single hypercharge boson. The  $L_a$ 's are  $SU(3)_C$  generators (the  $3 \times 3$  Gell-Mann matrices  $\frac{1}{2}\lambda_a$  for triplets, 0 for singlets), the  $T_b$ 's are  $SU(2)_L$  generators (the  $2 \times 2$  Pauli matrices  $\frac{1}{2}\tau_b$  for doublets, 0 for singlets), and the  $Y$ 's are the  $U(1)_Y$  charges. For example, for the quark doublets  $Q_L$ , we have

$$\mathcal{L}_{\text{kinetic}}(Q_L) = i\overline{Q_{Li}}\gamma_\mu \left( \partial^\mu + \frac{i}{2}g_s G_a^\mu \lambda_a + \frac{i}{2}g W_b^\mu \tau_b + \frac{i}{6}g' B^\mu \right) \delta_{ij} Q_{Lj}, \quad (6)$$

while for the lepton doublets  $L_L^I$ , we have

$$\mathcal{L}_{\text{kinetic}}(L_L) = i\overline{L_{Li}}\gamma_\mu \left( \partial^\mu + \frac{i}{2}g W_b^\mu \tau_b - \frac{i}{2}g' B^\mu \right) \delta_{ij} L_{Lj}. \quad (7)$$

The unit matrix in flavor space,  $\delta_{ij}$ , signifies that these parts of the interaction Lagrangian are flavor-universal. In addition, they conserve CP.

The Higgs potential, which describes the scalar self interactions, is given by:

$$\mathcal{L}_{\text{Higgs}} = \mu^2 \phi^\dagger \phi - \lambda (\phi^\dagger \phi)^2. \quad (8)$$

For the Standard Model scalar sector, where there is a single doublet, this part of the Lagrangian is also CP conserving.

The quark Yukawa interactions are given by

$$-\mathcal{L}_Y^q = Y_{ij}^d \overline{Q_{Li}} \phi D_{Rj} + Y_{ij}^u \overline{Q_{Li}} \tilde{\phi} U_{Rj} + \text{h.c.}, \quad (9)$$

(where  $\tilde{\phi} = i\tau_2 \phi^\dagger$ ) while the lepton Yukawa interactions are given by

$$-\mathcal{L}_Y^\ell = Y_{ij}^e \overline{L_{Li}} \phi E_{Rj} + \text{h.c.}. \quad (10)$$

This part of the Lagrangian is, in general, flavor-dependent (that is,  $Y^f \not\propto \mathbf{1}$ ) and CP violating.

### 3.2 Global symmetries

In the absence of the Yukawa matrices  $Y^d$ ,  $Y^u$  and  $Y^e$ , the SM has a large  $U(3)^5$  global symmetry:

$$G_{\text{global}}(Y^{u,d,e} = 0) = SU(3)_q^3 \times SU(3)_\ell^2 \times U(1)^5, \quad (11)$$

where

$$\begin{aligned} SU(3)_q^3 &= SU(3)_Q \times SU(3)_U \times SU(3)_D, \\ SU(3)_\ell^2 &= SU(3)_L \times SU(3)_E, \\ U(1)^5 &= U(1)_B \times U(1)_L \times U(1)_Y \times U(1)_{\text{PQ}} \times U(1)_E. \end{aligned} \quad (12)$$

Out of the five  $U(1)$  charges, three can be identified with baryon number ( $B$ ), lepton number ( $L$ ) and hypercharge ( $Y$ ), which are respected by the Yukawa interactions. The two remaining  $U(1)$  groups can be identified with the PQ symmetry whereby the Higgs and  $D_R, E_R$  fields have opposite charges, and with a global rotation of  $E_R$  only.

The point that is important for our purposes is that  $\mathcal{L}_{\text{kinetic}} + \mathcal{L}_{\text{Higgs}}$  respect the non-Abelian flavor symmetry  $S(3)_q^3 \times SU(3)_\ell^2$ , under which

$$Q_L \rightarrow V_Q Q_L, \quad U_R \rightarrow V_U U_R, \quad D_R \rightarrow V_D D_R, \quad L_L \rightarrow V_L L_L, \quad E_R \rightarrow V_E E_R, \quad (13)$$

where the  $V_i$  are unitary matrices. The Yukawa interactions (9) and (10) break the global symmetry,

$$G_{\text{global}}(Y^{u,d,e} \neq 0) = U(1)_B \times U(1)_e \times U(1)_\mu \times U(1)_\tau. \quad (14)$$

(Of course, the gauged  $U(1)_Y$  also remains a good symmetry.) Thus, the transformations of Eq. (13) are not a symmetry of  $\mathcal{L}_{\text{SM}}$ . Instead, they correspond to a change of the interaction basis. These observations also offer an alternative way of defining flavor physics: it refers to interactions that break the  $SU(3)^5$  symmetry (13). Thus, the term “**flavor violation**” is often used to describe processes or parameters that break the symmetry.

One can think of the quark Yukawa couplings as spurions that break the global  $SU(3)_q^3$  symmetry (but are neutral under  $U(1)_B$ ),

$$Y^u \sim (3, \bar{3}, 1)_{SU(3)_q^3}, \quad Y^d \sim (3, 1, \bar{3})_{SU(3)_q^3}, \quad (15)$$

and of the lepton Yukawa couplings as spurions that break the global  $SU(3)_\ell^2$  symmetry (but are neutral under  $U(1)_e \times U(1)_\mu \times U(1)_\tau$ ),

$$Y^e \sim (3, \bar{3})_{SU(3)_\ell^2}. \quad (16)$$

The spurion formalism is convenient for several purposes: parameter counting (see below), identification of flavor suppression factors (see Section 5), and the idea of minimal flavor violation (see Section 5.3).

### 3.3 Counting parameters

How many independent parameters are there in  $\mathcal{L}_Y^q$ ? The two Yukawa matrices,  $Y^u$  and  $Y^d$ , are  $3 \times 3$  and complex. Consequently, there are 18 real and 18 imaginary parameters in these matrices. Not all of them are, however, physical. The pattern of  $G_{\text{global}}$  breaking means that there is freedom to remove 9 real and 17 imaginary parameters (the number of parameters in three  $3 \times 3$  unitary matrices minus the phase related to  $U(1)_B$ ). For example, we can use the unitary transformations  $Q_L \rightarrow V_Q Q_L$ ,  $U_R \rightarrow V_U U_R$  and  $D_R \rightarrow V_D D_R$ , to lead to the following interaction basis:

$$Y^d = \lambda_d, \quad Y^u = V^\dagger \lambda_u, \quad (17)$$

where  $\lambda_{d,u}$  are diagonal,

$$\lambda_d = \text{diag}(y_d, y_s, y_b), \quad \lambda_u = \text{diag}(y_u, y_c, y_t), \quad (18)$$

while  $V$  is a unitary matrix that depends on three real angles and one complex phase. We conclude that there are 10 quark flavor parameters: 9 real ones and a single phase. In the mass basis, we will identify the nine real parameters as six quark masses and three mixing angles, while the single phase is  $\delta_{\text{KM}}$ .

How many independent parameters are there in  $\mathcal{L}_Y^\ell$ ? The Yukawa matrix  $Y^e$  is  $3 \times 3$  and complex. Consequently, there are 9 real and 9 imaginary parameters in this matrix. There is, however, freedom to remove 6 real and 9 imaginary parameters (the number of parameters in two  $3 \times 3$  unitary matrices minus the phases related to  $U(1)^3$ ). For example, we can use the unitary transformations  $L_L \rightarrow V_L L_L$  and  $E_R \rightarrow V_E E_R$ , to lead to the following interaction basis:

$$Y^e = \lambda_e = \text{diag}(y_e, y_\mu, y_\tau). \quad (19)$$

We conclude that there are 3 real lepton flavor parameters. In the mass basis, we will identify these parameters as the three charged lepton masses. We must, however, modify the model when we take into account the evidence for neutrino masses.

### 3.4 The mass basis

Upon the replacement  $\mathcal{R}e(\phi^0) \rightarrow \frac{v+h^0}{\sqrt{2}}$ , the Yukawa interactions (9) give rise to the mass matrices

$$M_q = \frac{v}{\sqrt{2}} Y^q. \quad (20)$$

The mass basis corresponds, by definition, to diagonal mass matrices. We can always find unitary matrices  $V_{qL}$  and  $V_{qR}$  such that

$$V_{qL} M_q V_{qR}^\dagger = M_q^{\text{diag}} \equiv \frac{v}{\sqrt{2}} \lambda_q. \quad (21)$$

The four matrices  $V_{dL}$ ,  $V_{dR}$ ,  $V_{uL}$  and  $V_{uR}$  are then the ones required to transform to the mass basis. For example, if we start from the special basis (17), we have  $V_{dL} = V_{dR} = V_{uR} = \mathbf{1}$  and  $V_{uL} = V$ . The combination  $V_{uL} V_{dL}^\dagger$  is independent of the interaction basis from which we start this procedure.

We denote the left-handed quark mass eigenstates as  $U_L$  and  $D_L$ . The charged current interactions for quarks [that is the interactions of the charged  $SU(2)_L$  gauge bosons  $W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2)$ ], which in the interaction basis are described by (6), have a complicated form in the mass basis:

$$-\mathcal{L}_{W^\pm}^q = \frac{g}{\sqrt{2}} \overline{U_{Li}} \gamma^\mu V_{ij} D_{Lj} W_\mu^+ + \text{h.c.} \quad (22)$$

where  $V$  is the  $3 \times 3$  unitary matrix ( $VV^\dagger = V^\dagger V = \mathbf{1}$ ) that appeared in Eq. (17). For a general interaction basis,

$$V = V_{uL} V_{dL}^\dagger. \quad (23)$$

$V$  is the Cabibbo-Kobayashi-Maskawa (CKM) *mixing matrix* for quarks [1, 2]. As a result of the fact that  $V$  is not diagonal, the  $W^\pm$  gauge bosons couple to quark mass eigenstates of different generations. Within the Standard Model, this is the only source of *flavor changing* quark interactions.

**Exercise 1:** Prove that, in the absence of neutrino masses, there is no mixing in the lepton sector.

**Exercise 2:** Prove that there is no mixing in the  $Z$  couplings. (In the physics jargon, there are no *flavor changing neutral currents at tree level*.)

The detailed structure of the CKM matrix, its parametrization, and the constraints on its elements are described in Appendix A.

## 4 Testing CKM

Measurements of rates, mixing, and CP asymmetries in  $B$  decays in the two  $B$  factories, BaBar and Belle, and in the two Tevatron detectors, CDF and D0, signified a new era in our understanding of CP violation. The progress is both qualitative and quantitative. Various basic questions concerning CP and flavor violation have received, for the first time, answers based on experimental information. These questions include, for example,

- Is the Kobayashi-Maskawa mechanism at work (namely, is  $\delta_{\text{KM}} \neq 0$ )?
- Does the KM phase dominate the observed CP violation?

As a first step, one may assume the SM and test the overall consistency of the various measurements. However, the richness of data from the  $B$  factories allow us to go a step further and answer these questions model independently, namely allowing new physics to contribute to the relevant processes. We here explain the way in which this analysis proceeds.

#### 4.1 $S_{\psi K_S}$

The CP asymmetry in  $B \rightarrow \psi K_S$  decays plays a major role in testing the KM mechanism. Before we explain the test itself, we should understand why the theoretical interpretation of the asymmetry is exceptionally clean, and what are the theoretical parameters on which it depends, within and beyond the Standard Model.

The CP asymmetry in neutral meson decays into final CP eigenstates  $f_{CP}$  is defined as follows:

$$\mathcal{A}_{f_{CP}}(t) \equiv \frac{d\Gamma/dt[\overline{B}_{\text{phys}}^0(t) \rightarrow f_{CP}] - d\Gamma/dt[B_{\text{phys}}^0(t) \rightarrow f_{CP}]}{d\Gamma/dt[\overline{B}_{\text{phys}}^0(t) \rightarrow f_{CP}] + d\Gamma/dt[B_{\text{phys}}^0(t) \rightarrow f_{CP}]} . \quad (24)$$

A detailed evaluation of this asymmetry is given in Appendix B. It leads to the following form:

$$\begin{aligned} \mathcal{A}_{f_{CP}}(t) &= S_{f_{CP}} \sin(\Delta mt) - C_{f_{CP}} \cos(\Delta mt), \\ S_{f_{CP}} &\equiv \frac{2\mathcal{I}m(\lambda_{f_{CP}})}{1 + |\lambda_{f_{CP}}|^2}, \quad C_{f_{CP}} \equiv \frac{1 - |\lambda_{f_{CP}}|^2}{1 + |\lambda_{f_{CP}}|^2}, \end{aligned} \quad (25)$$

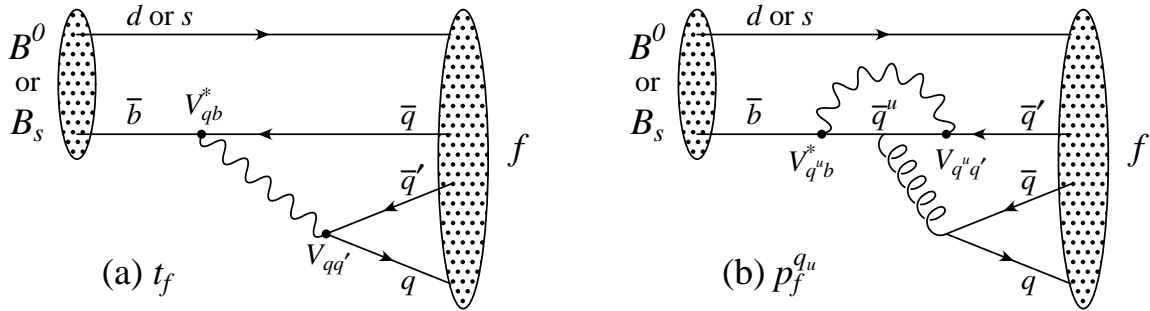
where

$$\lambda_{f_{CP}} = e^{-i\phi_B} (\overline{A}_{f_{CP}} / A_{f_{CP}}) . \quad (26)$$

Here  $\phi_B$  refers to the phase of  $M_{12}$  [see Eq. (B.23)]. Within the Standard Model, the corresponding phase factor is given by

$$e^{-i\phi_B} = (V_{tb}^* V_{td}) / (V_{tb} V_{td}^*) . \quad (27)$$

The decay amplitudes  $A_f$  and  $\overline{A}_f$  are defined in Eq. (B.1).



**Fig. 1:** Feynman diagrams for (a) tree and (b) penguin amplitudes contributing to  $B^0 \rightarrow f$  or  $B_s \rightarrow f$  via a  $\bar{b} \rightarrow \bar{q}q\bar{q}'$  quark-level process.

The  $B^0 \rightarrow J/\psi K^0$  decay [3,4] proceeds via the quark transition  $\bar{b} \rightarrow \bar{c}c\bar{s}$ . There are contributions from both tree ( $t$ ) and penguin ( $p^{qu}$ , where  $q_u = u, c, t$  is the quark in the loop) diagrams (see Fig. 1) which carry different weak phases:

$$A_f = (V_{cb}^* V_{cs}) t_f + \sum_{q_u=u,c,t} (V_{q_u b}^* V_{q_u s}) p_f^{q_u} . \quad (28)$$

(The distinction between tree and penguin contributions is a heuristic one, the separation by the operator that enters is more precise. For a detailed discussion of the more complete operator product approach, which also includes higher order QCD corrections, see, for example, ref. [5].) Using CKM unitarity, these decay amplitudes can always be written in terms of just two CKM combinations:

$$A_{\psi K} = (V_{cb}^* V_{cs}) T_{\psi K} + (V_{ub}^* V_{us}) P_{\psi K}^u, \quad (29)$$

where  $T_{\psi K} = t_{\psi K} + p_{\psi K}^c - p_{\psi K}^t$  and  $P_{\psi K}^u = p_{\psi K}^u - p_{\psi K}^t$ . A subtlety arises in this decay that is related to the fact that  $B^0 \rightarrow J/\psi K^0$  and  $\bar{B}^0 \rightarrow J/\psi \bar{K}^0$ . A common final state, e.g.  $J/\psi K_S$ , can be reached via  $K^0 - \bar{K}^0$  mixing. Consequently, the phase factor corresponding to neutral  $K$  mixing,  $e^{-i\phi_K} = (V_{cd}^* V_{cs}) / (V_{cb} V_{cs}^*)$ , plays a role:

$$\frac{\bar{A}_{\psi K_S}}{A_{\psi K_S}} = -\frac{(V_{cb} V_{cs}^*) T_{\psi K} + (V_{ub} V_{us}^*) P_{\psi K}^u}{(V_{cb}^* V_{cs}) T_{\psi K} + (V_{ub}^* V_{us}) P_{\psi K}^u} \times \frac{V_{cd}^* V_{cs}}{V_{cb} V_{cs}^*}. \quad (30)$$

The crucial point is that, for  $B \rightarrow J/\psi K_S$  and other  $\bar{b} \rightarrow \bar{c} c \bar{s}$  processes, we can neglect the  $P^u$  contribution to  $A_{\psi K}$ , in the SM, to an approximation that is better than one percent:

$$|P_{\psi K}^u / T_{\psi K}| \times |V_{ub} / V_{cb}| \times |V_{us} / V_{cs}| \sim (\text{loop factor}) \times 0.1 \times 0.23 \lesssim 0.005. \quad (31)$$

Thus, to an accuracy better than one percent,

$$\lambda_{\psi K_S} = \left( \frac{V_{tb}^* V_{td}}{V_{tb} V_{td}^*} \right) \left( \frac{V_{cb} V_{cd}^*}{V_{cb}^* V_{cd}} \right) = -e^{-2i\beta}, \quad (32)$$

where  $\beta$  is defined in Eq. (A.9), and consequently

$$S_{\psi K_S} = \sin 2\beta, \quad C_{\psi K_S} = 0. \quad (33)$$

(Below the percent level, several effects modify this equation [6–9].)

**Exercise 3:** Show that, if the  $B \rightarrow \pi\pi$  decays were dominated by tree diagrams, then  $S_{\pi\pi} = \sin 2\alpha$ .

**Exercise 4:** Estimate the accuracy of the predictions  $S_{\phi K_S} = \sin 2\beta$  and  $C_{\phi K_S} = 0$ .

When we consider extensions of the SM, we still do not expect any significant new contribution to the tree level decay,  $b \rightarrow c \bar{c} s$ , beyond the SM  $W$ -mediated diagram. Thus, the expression  $\bar{A}_{\psi K_S} / A_{\psi K_S} = (V_{cb} V_{cd}^*) / (V_{cb}^* V_{cd})$  remains valid, though the approximation of neglecting sub-dominant phases can be somewhat less accurate than Eq. (31). On the other hand,  $M_{12}$ , the  $B^0 - \bar{B}^0$  mixing amplitude, can in principle get large and even dominant contributions from new physics. We can parametrize the modification to the SM in terms of two parameters,  $r_d^2$  signifying the change in magnitude, and  $2\theta_d$  signifying the change in phase:

$$M_{12} = r_d^2 e^{2i\theta_d} M_{12}^{\text{SM}}(\rho, \eta). \quad (34)$$

This leads to the following generalization of Eq. (33):

$$S_{\psi K_S} = \sin(2\beta + 2\theta_d), \quad C_{\psi K_S} = 0. \quad (35)$$

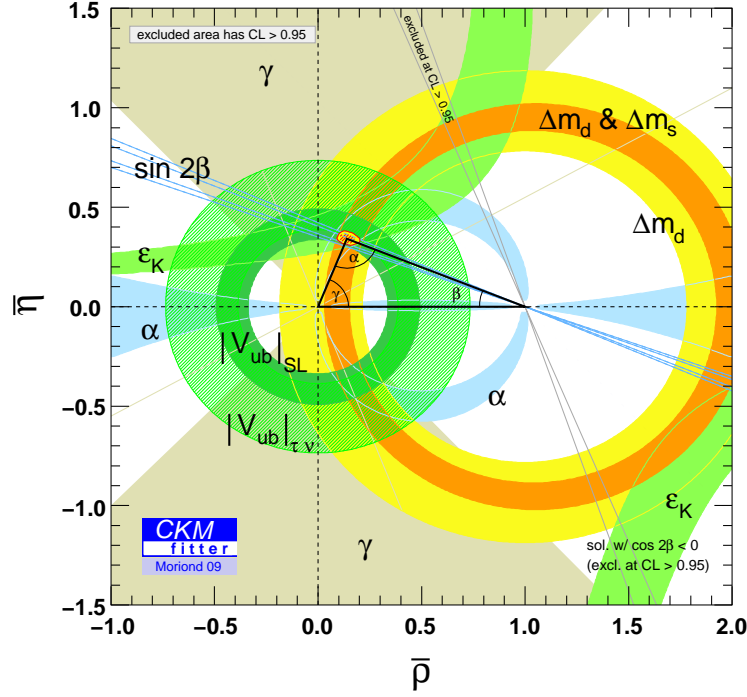
The experimental measurements give the following ranges [10]:

$$S_{\psi K_S} = +0.68 \pm 0.02, \quad C_{\psi K_S} = +0.005 \pm 0.017. \quad (36)$$

## 4.2 Self-consistency of the CKM assumption

The three generation standard model has room for CP violation, through the KM phase in the quark mixing matrix. Yet, one would like to make sure that indeed CP is violated by the SM interactions, namely that  $\sin \delta_{\text{KM}} \neq 0$ . If we establish that this is the case, we would further like to know whether the SM contributions to CP violating observables are dominant. More quantitatively, we would like to put an upper bound on the ratio between the new physics and the SM contributions.

As a first step, one can assume that flavor changing processes are fully described by the SM, and check the consistency of the various measurements with this assumption. There are four relevant mixing



**Fig. 2:** Allowed region in the  $\rho, \eta$  plane. Superimposed are the individual constraints from charmless semileptonic  $B$  decays ( $|V_{ub}/V_{cb}|$ ), mass differences in the  $B^0$  ( $\Delta m_d$ ) and  $B_s$  ( $\Delta m_s$ ) neutral meson systems, and CP violation in  $K \rightarrow \pi\pi$  ( $\epsilon_K$ ),  $B \rightarrow \psi K$  ( $\sin 2\beta$ ),  $B \rightarrow \pi\pi, \rho\pi, \rho\rho$  ( $\alpha$ ), and  $B \rightarrow DK$  ( $\gamma$ ). Taken from [12].

parameters, which can be taken to be the Wolfenstein parameters  $\lambda, A, \rho$  and  $\eta$  defined in Eq. (A.4). The values of  $\lambda$  and  $A$  are known rather accurately [11] from, respectively,  $K \rightarrow \pi\ell\nu$  and  $b \rightarrow c\ell\nu$  decays:

$$\lambda = 0.2254 \pm 0.0007, \quad A = 0.811^{+0.022}_{-0.012}. \quad (37)$$

Then, one can express all the relevant observables as a function of the two remaining parameters,  $\rho$  and  $\eta$ , and check whether there is a range in the  $\rho - \eta$  plane that is consistent with all measurements. The list of observables includes the following:

- The rates of inclusive and exclusive charmless semileptonic  $B$  decays depend on  $|V_{ub}|^2 \propto \rho^2 + \eta^2$ ;
- The CP asymmetry in  $B \rightarrow \psi K_S, S_{\psi K_S} = \sin 2\beta = \frac{2\eta(1-\rho)}{(1-\rho)^2 + \eta^2}$ ;
- The rates of various  $B \rightarrow DK$  decays depend on the phase  $\gamma$ , where  $e^{i\gamma} = \frac{\rho+i\eta}{\sqrt{\rho^2 + \eta^2}}$ ;
- The rates of various  $B \rightarrow \pi\pi, \rho\pi, \rho\rho$  decays depend on the phase  $\alpha = \pi - \beta - \gamma$ ;
- The ratio between the mass splittings in the neutral  $B$  and  $B_s$  systems is sensitive to  $|V_{td}/V_{ts}|^2 = \lambda^2[(1-\rho)^2 + \eta^2]$ ;
- The CP violation in  $K \rightarrow \pi\pi$  decays,  $\epsilon_K$ , depends in a complicated way on  $\rho$  and  $\eta$ .

The resulting constraints are shown in Fig. 2.

The consistency of the various constraints is impressive. In particular, the following ranges for  $\rho$  and  $\eta$  can account for all the measurements [11]:

$$\rho = +0.131^{+0.026}_{-0.013}, \quad \eta = +0.345 \pm 0.014. \quad (38)$$

One can make then the following statement [13]:

**Very likely, CP violation in flavor changing processes is dominated by the Kobayashi-Maskawa phase.**

In the next two subsections, we explain how we can remove the phrase “very likely” from this statement, and how we can quantify the KM-dominance.

### 4.3 Is the KM mechanism at work?

In proving that the KM mechanism is at work, we assume that charged-current tree-level processes are dominated by the  $W$ -mediated SM diagrams (see, for example, [14]). This is a very plausible assumption. I am not aware of any viable well-motivated model where this assumption is not valid. Thus we can use all tree level processes and fit them to  $\rho$  and  $\eta$ , as we did before. The list of such processes includes the following:

1. Charmless semileptonic  $B$ -decays,  $b \rightarrow u\ell\nu$ , measure  $R_u$  [see Eq. (A.8)].
2.  $B \rightarrow DK$  decays, which go through the quark transitions  $b \rightarrow c\bar{u}s$  and  $b \rightarrow u\bar{c}s$ , measure the angle  $\gamma$  [see Eq. (A.9)].
3.  $B \rightarrow \rho\rho$  decays (and, similarly,  $B \rightarrow \pi\pi$  and  $B \rightarrow \rho\pi$  decays) go through the quark transition  $b \rightarrow u\bar{u}d$ . With an isospin analysis, one can determine the relative phase between the tree decay amplitude and the mixing amplitude. By incorporating the measurement of  $S_{\psi K_S}$ , one can subtract the phase from the mixing amplitude, finally providing a measurement of the angle  $\gamma$  [see Eq. (A.9)].

In addition, we can use loop processes, but then we must allow for new physics contributions, in addition to the  $(\rho, \eta)$ -dependent SM contributions. Of course, if each such measurement adds a separate mode-dependent parameter, then we do not gain anything by using this information. However, there is a number of observables where the only relevant loop process is  $B^0 - \bar{B}^0$  mixing. The list includes  $S_{\psi K_S}$ ,  $\Delta m_B$  and the CP asymmetry in semileptonic  $B$  decays:

$$\begin{aligned} S_{\psi K_S} &= \sin(2\beta + 2\theta_d), \\ \Delta m_B &= r_d^2 (\Delta m_B)^{\text{SM}}, \\ \mathcal{A}_{\text{SL}} &= -\mathcal{R}e \left( \frac{\Gamma_{12}}{M_{12}} \right)^{\text{SM}} \frac{\sin 2\theta_d}{r_d^2} + \mathcal{I}m \left( \frac{\Gamma_{12}}{M_{12}} \right)^{\text{SM}} \frac{\cos 2\theta_d}{r_d^2}. \end{aligned} \quad (39)$$

As explained above, such processes involve two new parameters [see Eq. (34)]. Since there are three relevant observables, we can further tighten the constraints in the  $(\rho, \eta)$ -plane. Similarly, one can use measurements related to  $B_s - \bar{B}_s$  mixing. One gains three new observables at the cost of two new parameters (see, for example, [15]).

The results of such fit, projected on the  $\rho - \eta$  plane, can be seen in Fig. 3. It gives [12]

$$\eta = 0.44_{-0.23}^{+0.05} \quad (3\sigma). \quad (40)$$

[A similar analysis in Ref. [16] obtains the  $3\sigma$  range  $(0.31 - 0.46)$ .] It is clear that  $\eta \neq 0$  is well established:

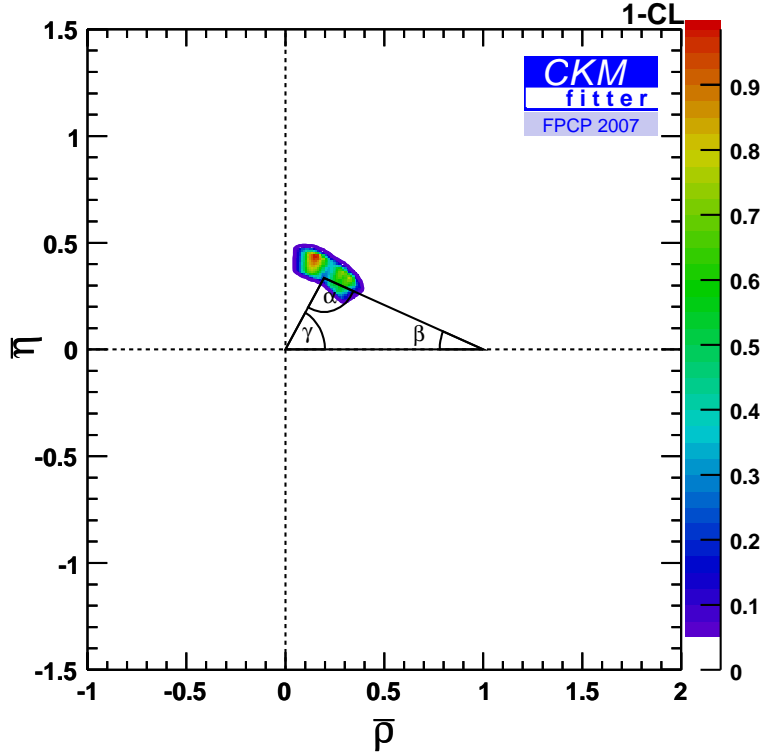
**The Kobayashi-Maskawa mechanism of CP violation is at work.**

Another way to establish that CP is violated by the CKM matrix is to find, within the same procedure, the allowed range for  $\sin 2\beta$  [16]:

$$\sin 2\beta^{\text{tree}} = 0.80 \pm 0.03. \quad (41)$$

Thus,  $\beta \neq 0$  is well established.

The consistency of the experimental results (36) with the SM predictions (33,41) means that the KM mechanism of CP violation dominates the observed CP violation. In the next subsection, we make this statement more quantitative.



**Fig. 3:** The allowed region in the  $\rho - \eta$  plane, assuming that tree diagrams are dominated by the Standard Model [12].

#### 4.4 How much can new physics contribute to $B^0 - \bar{B}^0$ mixing?

All that we need to do in order to establish whether the SM dominates the observed CP violation, and to put an upper bound on the new physics contribution to  $B^0 - \bar{B}^0$  mixing, is to project the results of the fit performed in the previous subsection on the  $r_d^2 - 2\theta_d$  plane. If we find that  $\theta_d \ll \beta$ , then the SM dominance in the observed CP violation will be established. The constraints are shown in Fig. 4(a). Indeed,  $\theta_d \ll \beta$ .

An alternative way to present the data is to use the  $h_d, \sigma_d$  parametrization,

$$r_d^2 e^{2i\theta_d} = 1 + h_d e^{2i\sigma_d}. \quad (42)$$

While the  $r_d, \theta_d$  parameters give the relation between the full mixing amplitude and the SM one, and are convenient to apply to the measurements, the  $h_d, \sigma_d$  parameters give the relation between the new physics and SM contributions, and are more convenient in testing theoretical models:

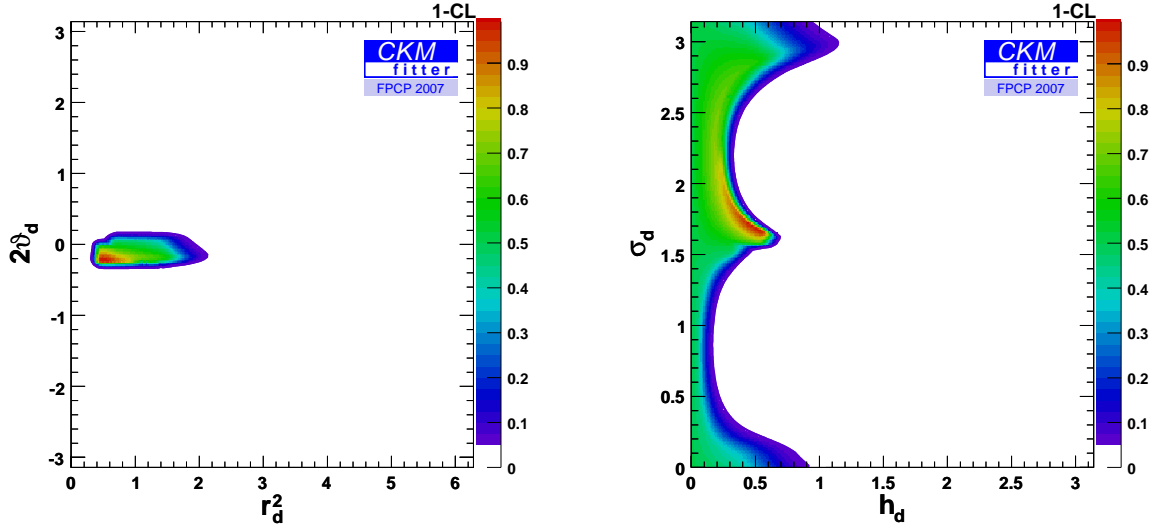
$$h_d e^{2i\sigma_d} = \frac{M_{12}^{\text{NP}}}{M_{12}^{\text{SM}}}. \quad (43)$$

The constraints in the  $h_d - \sigma_d$  plane are shown in Fig. 4(b). We can make the following two statements:

1. A new physics contribution to  $B^0 - \bar{B}^0$  mixing amplitude that carries a phase that is significantly different from the KM phase is constrained to lie below the 20-30% level.
2. A new physics contribution to the  $B^0 - \bar{B}^0$  mixing amplitude which is aligned with the KM phase is constrained to be at most comparable to the CKM contribution.

One can reformulate these statements as follows:





**Fig. 4:** Constraints in the (a)  $r_d^2 - 2\theta_d$  plane, and (b)  $h_d - \sigma_d$  plane, assuming that NP contributions to tree level processes are negligible [12].

1. The KM mechanism dominates CP violation in  $B^0 - \bar{B}^0$  mixing.
2. The CKM mechanism is a major player in  $B^0 - \bar{B}^0$  mixing.

## 5 The new physics flavor puzzle

### 5.1 A model independent discussion

It is clear that the Standard Model is not a complete theory of Nature:

1. It does not include gravity, and therefore it cannot be valid at energy scales above  $m_{\text{Planck}} \sim 10^{19}$  GeV;
2. It does not allow for neutrino masses, and therefore it cannot be valid at energy scales above  $m_{\text{seesaw}} \sim 10^{15}$  GeV;
3. The fine-tuning problem of the Higgs mass suggests that the scale where the SM is replaced with a more fundamental theory is actually much lower,  $m_{\text{top-partners}} \lesssim$  a few TeV.
4. If the dark matter is made of weakly interacting massive particles (WIMPs) then, again, a low scale of new physics is likely,  $m_{\text{wimp}} \lesssim$  a few TeV.

Given that the SM is only an effective low energy theory, non-renormalizable terms must be added to  $\mathcal{L}_{\text{SM}}$  of Eq. (4). These are terms of dimension higher than four in the fields which, therefore, have couplings that are inversely proportional to the scale of new physics  $\Lambda_{\text{NP}}$ . For example, the lowest dimension non-renormalizable terms are dimension five:

$$-\mathcal{L}_{\text{Yukawa}}^{\text{dim-5}} = \frac{Z_{ij}^\nu}{\Lambda_{\text{NP}}} L_{Li}^I L_{Lj}^I \phi \phi + \text{h.c.} \quad (44)$$

These are the seesaw terms, leading to neutrino masses.

**Exercise 5:** How does the global symmetry breaking pattern (14) change when (44) is taken into account?

**Exercise 6:** What is the number of physical lepton flavor parameters in this case? Identify these parameters in the mass basis.

**Table 1:** Measurements related to neutral meson mixing

Sector	CP-conserving	CP-violating
sd	$\Delta m_K/m_K = 7.0 \times 10^{-15}$	$\epsilon_K = 2.3 \times 10^{-3}$
cu	$\Delta m_D/m_D = 8.7 \times 10^{-15}$	$A_\Gamma/y_{\text{CP}} \lesssim 0.2$
bd	$\Delta m_B/m_B = 6.3 \times 10^{-14}$	$S_{\psi K} = +0.67 \pm 0.02$
bs	$\Delta m_{B_s}/m_{B_s} = 2.1 \times 10^{-12}$	$S_{\psi\phi} = -0.04 \pm 0.09$

**Table 2:** Lower bounds on the scale of new physics  $\Lambda_{\text{NP}}$ , in units of TeV. The bounds from CP conserving (violating) observables scale like  $\sqrt{z_{ij}}$  ( $\sqrt{z_{ij}^I}$ ).

$ij$	CP-conserving	CP-violating
sd	$1 \times 10^3$	$2 \times 10^4$
cu	$1 \times 10^3$	$3 \times 10^3$
bd	$4 \times 10^2$	$8 \times 10^2$
bs	$7 \times 10^1$	$2 \times 10^2$

As concerns quark flavor physics, consider, for example, the following dimension-six, four-fermion, flavor changing operators:

$$\mathcal{L}_{\Delta F=2} = \frac{z_{sd}}{\Lambda_{\text{NP}}^2} (\overline{d_L} \gamma_\mu s_L)^2 + \frac{z_{cu}}{\Lambda_{\text{NP}}^2} (\overline{c_L} \gamma_\mu u_L)^2 + \frac{z_{bd}}{\Lambda_{\text{NP}}^2} (\overline{d_L} \gamma_\mu b_L)^2 + \frac{z_{bs}}{\Lambda_{\text{NP}}^2} (\overline{s_L} \gamma_\mu b_L)^2. \quad (45)$$

Each of these terms contributes to the mass splitting between the corresponding two neutral mesons. For example, the term  $\mathcal{L}_{\Delta B=2} \propto (\overline{d_L} \gamma_\mu b_L)^2$  contributes to  $\Delta m_B$ , the mass difference between the two neutral  $B$ -mesons. We use  $M_{12}^B = \frac{1}{2m_B} \langle B^0 | \mathcal{L}_{\Delta F=2} | \overline{B}^0 \rangle$  and

$$\langle B^0 | (\overline{d_{La}} \gamma^\mu b_{La}) (\overline{d_{Lb}} \gamma_\mu b_{Lb}) | \overline{B}^0 \rangle = -\frac{1}{3} m_B^2 f_B^2 B_B. \quad (46)$$

This leads to  $\Delta m_B/m_B = 2|M_{12}^B|/m_B \sim (|z_{bd}|/3)(f_B/\Lambda_{\text{NP}})^2$ . Analogous expressions hold for the other neutral mesons.

The experimental results for CP conserving and CP violating observables related to neutral meson mixing (mass splittings and CP asymmetries in tree level decays, respectively) are given in Table 1.

The measurements quoted in Table 1 lead, for a given value of  $|z_{ij}|$  and  $z_{ij}^I \equiv \mathcal{I}m(z_{ij})$ , to lower bounds on the scale  $\Lambda_{\text{NP}}$ . In Table 2 we give the bounds that correspond to  $|z_{ij}| = 1$  and to  $z_{ij}^I = 1$ . The bounds scale like  $\sqrt{z_{ij}}$  and  $\sqrt{z_{ij}^I}$ , respectively.

We conclude that if the new physics has a generic flavor structure, that is  $z_{ij} = \mathcal{O}(1)$ , then its scale must be above  $10^3 - 10^4$  TeV. If the leading contributions involve electroweak loops, the lower bound is somewhat lower, of order  $10^2 - 10^3$  TeV. The bounds from the corresponding four-fermi terms with LR structure, instead of the LL structure of Eq. (45), are even stronger. *If indeed  $\Lambda_{\text{NP}} \gg \text{TeV}$ , it means that we have misinterpreted the hints from the fine-tuning problem and the dark matter puzzle.*

There is, however, another way to look at these constraints:

$$z_{sd} \lesssim 8 \times 10^{-7} (\Lambda_{\text{NP}}/\text{TeV})^2,$$

$$\begin{aligned}
 z_{cu} &\lesssim 5 \times 10^{-7} (\Lambda_{\text{NP}}/TeV)^2, \\
 z_{bd} &\lesssim 5 \times 10^{-6} (\Lambda_{\text{NP}}/TeV)^2, \\
 z_{bs} &\lesssim 2 \times 10^{-4} (\Lambda_{\text{NP}}/TeV)^2,
 \end{aligned} \tag{47}$$

$$\begin{aligned}
 z_{sd}^I &\lesssim 6 \times 10^{-9} (\Lambda_{\text{NP}}/TeV)^2, \\
 z_{cu}^I &\lesssim 1 \times 10^{-7} (\Lambda_{\text{NP}}/TeV)^2, \\
 z_{bd}^I &\lesssim 1 \times 10^{-6} (\Lambda_{\text{NP}}/TeV)^2, \\
 z_{bs}^I &\lesssim 2 \times 10^{-5} (\Lambda_{\text{NP}}/TeV)^2.
 \end{aligned} \tag{48}$$

It could be that the scale of new physics is of order TeV, but its flavor structure is far from generic. Specifically, if new particles at the TeV scale couple to the SM fermions, then there are two ways in which their contributions to FCNC processes, such as neutral meson mixing, can be suppressed: degeneracy and alignment. Either of these principles, or a combination of both, signifies non-generic structure.

One can use the language of effective operators also for the SM, integrating out all particles significantly heavier than the neutral mesons (that is, the top, the Higgs and the weak gauge bosons). Thus, the scale is  $\Lambda_{\text{SM}} \sim m_W$ . Since the leading contributions to neutral meson mixings come from box diagrams, the  $z_{ij}$  coefficients are suppressed by  $\alpha_2^2$ . To identify the relevant flavor suppression factor, one can employ the spurion formalism. For example, the flavor transition that is relevant to  $B^0 - \bar{B}^0$  mixing involves  $\bar{d}_L b_L$  which transforms as  $(8, 1, 1)_{SU(3)_q}$ . The leading contribution must then be proportional to  $(Y^u Y^{u\dagger})_{13} \propto y_t^2 V_{tb} V_{td}^*$ . Indeed, an explicit calculation, using VIA for the matrix element and neglecting QCD corrections, gives (a detailed derivation can be found in Appendix B of [17])

$$\frac{2M_{12}^B}{m_B} \approx -\frac{\alpha_2^2}{12} \frac{f_B^2}{m_W^2} S_0(x_t) (V_{tb} V_{td}^*)^2, \tag{49}$$

where  $x_i = m_i^2/m_W^2$  and

$$S_0(x) = \frac{x}{(1-x)^2} \left[ 1 - \frac{11x}{4} + \frac{x^2}{4} - \frac{3x^2 \ln x}{2(1-x)} \right]. \tag{50}$$

Similar spurion analyses, or explicit calculations, allow us to extract the weak and flavor suppression factors that apply in the SM:

$$\begin{aligned}
 \mathcal{I}m(z_{sd}^{\text{SM}}) &\sim \alpha_2^2 y_t^2 |V_{td} V_{ts}|^2 \sim 1 \times 10^{-10}, \\
 z_{sd}^{\text{SM}} &\sim \alpha_2^2 y_c^2 |V_{cd} V_{cs}|^2 \sim 5 \times 10^{-9}, \\
 \mathcal{I}m(z_{cu}^{\text{SM}}) &\sim \alpha_2^2 y_b^2 |V_{ub} V_{cb}|^2 \sim 2 \times 10^{-14}, \\
 z_{bd}^{\text{SM}} &\sim \alpha_2^2 y_t^2 |V_{td} V_{tb}|^2 \sim 7 \times 10^{-8}, \\
 z_{bs}^{\text{SM}} &\sim \alpha_2^2 y_t^2 |V_{ts} V_{tb}|^2 \sim 2 \times 10^{-6}.
 \end{aligned} \tag{51}$$

Note that we did not include  $z_{cu}^{\text{SM}}$  in the list. The reason is that it requires a more detailed consideration. The naively leading short distance contribution is  $\propto \alpha_2^2 (y_s^4/y_c^2) |V_{cs} V_{us}|^2 \sim 5 \times 10^{-13}$ . However, higher dimension terms can replace a  $y_s^2$  factor with  $(\Lambda/m_D)^2$  [18]. Moreover, long distance contributions are expected to dominate. In particular, peculiar phase space effects [19, 20] have been identified which are expected to enhance  $\Delta m_D$  to within an order of magnitude of its measured value. The CP violating part, on the other hand, is dominated by short distance physics.

It is clear then that contributions from new physics at  $\Lambda_{\text{NP}} \sim 1 \text{ TeV}$  should be suppressed by factors that are comparable or smaller than the SM ones. Why does that happen? This is the new physics flavor puzzle.

**Table 3:** The phenomenological upper bounds on  $(\delta_{LL}^q)_{ij}$  and  $\langle \delta_{ij}^q \rangle = \sqrt{(\delta_{LL}^q)_{ij}(\delta_{RR}^q)_{ij}}$ . Here  $q = u, d$  and  $M = L, R$ . The constraints are given for  $m_{\tilde{q}} = 1$  TeV and  $x = m_{\tilde{g}}^2/m_{\tilde{q}}^2 = 1$ . We assume that the phases could suppress the imaginary part by a factor of  $\sim 0.3$ . Taken from Ref. [22].

$q$	$ij$	$(\delta_{LL}^q)_{ij}$	$\langle \delta_{ij}^q \rangle$
d	12	0.03	0.002
d	13	0.2	0.07
d	23	0.2	0.07
u	12	0.1	0.008

The fact that the flavor structure of new physics at the TeV scale must be non-generic means that flavor measurements are a good probe of the new physics. Perhaps the best-studied example is that of supersymmetry. Here, the spectrum of the superpartners and the structure of their couplings to the SM fermions will allow us to probe the mechanism of dynamical supersymmetry breaking.

## 5.2 The supersymmetric flavor puzzle

We consider, as an example, the contributions from the box diagrams involving the squark doublets of the second and third generations,  $\tilde{Q}_{L2,3}$ , to the  $B_s - \bar{B}_s$  mixing amplitude. The contributions are proportional to  $K_{3i}^{d*} K_{2i}^d K_{3j}^{d*} K_{2j}^d$ , where  $K^d$  is the mixing matrix of the gluino couplings to a left-handed down quark and their supersymmetric squark partners ( $\propto [(\delta_{LL}^d)_{23}]^2$  in the mass insertion approximation, described in Appendix C.1). We work in the mass basis for both quarks and squarks. A detailed derivation [21] is given in Appendix C.2. It gives:

$$M_{12}^s = \frac{\alpha_s^2 m_{B_s} f_{B_s}^2 B_{B_s} \eta_{\text{QCD}}}{108 m_{\tilde{d}}^2} [11 \tilde{f}_6(x) + 4x f_6(x)] \frac{(\Delta \tilde{m}_{\tilde{d}}^2)^2}{\tilde{m}_{\tilde{d}}^4} (K_{32}^{d*} K_{22}^d)^2. \quad (52)$$

Here  $m_{\tilde{d}}$  is the average mass of the two squark generations,  $\Delta m_{\tilde{d}}^2$  is the mass-squared difference, and  $x = m_{\tilde{g}}^2/m_{\tilde{d}}^2$ .

Eq. (52) can be translated into our generic language:

$$\Lambda_{\text{NP}} = m_{\tilde{q}}, \quad (53)$$

$$z_1^{bs} = \frac{11 \tilde{f}_6(x) + 4x f_6(x)}{18} \alpha_s^2 \left( \frac{\Delta \tilde{m}_{\tilde{d}}^2}{m_{\tilde{d}}^2} \right)^2 (K_{32}^{d*} K_{22}^d)^2 \approx 10^{-4} (\delta_{23}^{LL})^2,$$

where, for the last approximation, we took the example of  $x = 1$  [and used, correspondingly,  $11 \tilde{f}_6(1) + 4f_6(1) = 1/6$ ], and defined

$$\delta_{23}^{LL} = \left( \frac{\Delta \tilde{m}_{\tilde{d}}^2}{m_{\tilde{d}}^2} \right) (K_{32}^{d*} K_{22}^d). \quad (54)$$

Similar expressions can be derived for the dependence of  $K^0 - \bar{K}^0$  on  $(\delta_{MN}^d)_{12}$ ,  $B^0 - \bar{B}^0$  on  $(\delta_{MN}^d)_{13}$ , and  $D^0 - \bar{D}^0$  on  $(\delta_{MN}^u)_{12}$ . Then we can use the constraints of Eqs. (47,48) to put upper bounds on  $(\delta_{MN}^q)_{ij}$ . Some examples are given in Table 3 (see Ref. [22] for details and list of references).

We learn that, in most cases, we need  $\delta_{ij}^q/m_{\tilde{q}} \ll 1/\text{TeV}$ . One can immediately identify three generic ways in which supersymmetric contributions to neutral meson mixing can be suppressed:

1. Heaviness:  $m_{\tilde{q}} \gg 1 \text{ TeV}$ ;

2. Degeneracy:  $\Delta m_{\tilde{q}}^2 \ll m_{\tilde{q}}^2$ ;
3. Alignment:  $K_{ij}^q \ll 1$ .

When heaviness is the only suppression mechanism, as in split supersymmetry [23], the squarks are very heavy and supersymmetry no longer solves the fine tuning problem. (When the first two squark generations are mildly heavy and the third generation is light, as in effective supersymmetry [24], the fine tuning problem is still solved, but additional suppression mechanisms are needed.) If we want to maintain supersymmetry as a solution to the fine tuning problem, either degeneracy or alignment or a combination of the two is needed. This means that the flavor structure of supersymmetry is not generic, as argued in the previous section.

Take, for example,  $(\delta_{LL}^d)_{12} \leq 0.03$ . Naively, one might expect the alignment to be of order  $(V_{cd}V_{cs}^*) \sim 0.2$ , which is far from sufficient by itself. Barring a very precise alignment ( $|K_{12}^d| \ll |V_{us}|$ ) [25, 26] and accidental cancelations, we are led to conclude that the first two squark generations must be quasi-degenerate. Actually, by combining the constraints from  $K^0 - \bar{K}^0$  mixing and  $D^0 - \bar{D}^0$  mixing, one can show that this is the case independently of assumptions about the alignment [27–29]. Analogous conclusions can be drawn for many TeV-scale new physics scenarios: a strong level of degeneracy is required (for definitions and detailed analysis, see [30]).

**Exercise 9:** Does  $K_{31}^d \sim |V_{ub}|$  suffice to satisfy the  $\Delta m_B$  constraint with neither degeneracy nor heaviness? (Use the two generation approximation and ignore the second generation.)

Is there a natural way to make the squarks degenerate? Degeneracy requires that the  $3 \times 3$  matrix of soft supersymmetry breaking mass-squared terms  $\tilde{m}_{\tilde{Q}_L}^2 \simeq \tilde{m}_{\tilde{q}}^2 \mathbf{1}$ . We have mentioned already that flavor universality is a generic feature of gauge interactions. Thus, the requirement of degeneracy is perhaps a hint that supersymmetry breaking is *gauge mediated* to the MSSM fields.

### 5.3 Minimal flavor violation (MFV)

If supersymmetry breaking is gauge mediated, the squark mass matrices for  $SU(2)_L$ -doublet and  $SU(2)_L$ -singlet squarks have the following form at the scale of mediation  $m_M$ :

$$\begin{aligned}
 \tilde{M}_{\tilde{U}_L}^2(m_M) &= \left(m_{\tilde{Q}_L}^2 + D_{U_L}\right) \mathbf{1} + M_u M_u^\dagger, \\
 \tilde{M}_{\tilde{D}_L}^2(m_M) &= \left(m_{\tilde{Q}_L}^2 + D_{D_L}\right) \mathbf{1} + M_d M_d^\dagger, \\
 \tilde{M}_{\tilde{U}_R}^2(m_M) &= \left(m_{\tilde{U}_R}^2 + D_{U_R}\right) \mathbf{1} + M_u^\dagger M_u, \\
 \tilde{M}_{\tilde{D}_R}^2(m_M) &= \left(m_{\tilde{D}_R}^2 + D_{D_R}\right) \mathbf{1} + M_d^\dagger M_d,
 \end{aligned} \tag{55}$$

where  $D_{q_A} = (T_3)_{q_A} - (Q_{EM})_{q_A} s_W^2 m_Z^2 \cos 2\beta$  are the  $D$ -term contributions. Here, the only source of the  $SU(3)_q^3$  breaking are the SM Yukawa matrices.

This statement holds also when the renormalization group evolution is applied to find the form of these matrices at the weak scale. Taking the scale of the soft breaking terms  $m_{\tilde{q}_A}$  to be somewhat higher than the electroweak breaking scale  $m_Z$  allows us to neglect the  $D_{q_A}$  and  $M_q$  terms in (55). Then we obtain

$$\begin{aligned}
 \tilde{M}_{\tilde{Q}_L}^2(m_Z) &\sim m_{\tilde{Q}_L}^2 \left(r_3 \mathbf{1} + c_u Y_u Y_u^\dagger + c_d Y_d Y_d^\dagger\right), \\
 \tilde{M}_{\tilde{U}_R}^2(m_Z) &\sim m_{\tilde{U}_R}^2 \left(r_3 \mathbf{1} + c_{uR} Y_u^\dagger Y_u\right), \\
 \tilde{M}_{\tilde{D}_R}^2(m_Z) &\sim m_{\tilde{D}_R}^2 \left(r_3 \mathbf{1} + c_{dR} Y_d^\dagger Y_d\right).
 \end{aligned} \tag{56}$$

Here  $r_3$  represents the universal RGE contribution that is proportional to the gluino mass ( $r_3 = \mathcal{O}(6) \times (M_3(m_M)/m_{\tilde{q}}(m_M))$ ) and the  $c$ -coefficients depend logarithmically on  $m_M/m_Z$  and can be of  $\mathcal{O}(1)$  when  $m_M$  is not far below the GUT scale.

Models of gauge mediated supersymmetry breaking (GMSB) provide a concrete example of a large class of models that obey a simple principle called *minimal flavor violation* (MFV) [31]. This principle guarantees that low energy flavor changing processes deviate only very little from the SM predictions. The basic idea can be described as follows. The gauge interactions of the SM are universal in flavor space. The only breaking of this flavor universality comes from the three Yukawa matrices,  $Y^u$ ,  $Y^d$  and  $Y^e$ . If this remains true in the presence of the new physics, namely  $Y^u$ ,  $Y^d$  and  $Y^e$  are the only flavor non-universal parameters, then the model belongs to the MFV class.

Let us now formulate this principle in a more formal way, using the language of spurions that we presented in section 3.2. The Standard Model with vanishing Yukawa couplings has a large global symmetry (11,12). In this section we concentrate only on the quarks. The non-Abelian part of the flavor symmetry for the quarks is  $SU(3)_q^3$  of Eq. (12) with the three generations of quark fields transforming as follows:

$$Q_L(3, 1, 1), \quad U_R(1, 3, 1), \quad D_R(1, 1, 3). \quad (57)$$

The Yukawa interactions,

$$\mathcal{L}_Y = \overline{Q}_L Y^d D_R H + \overline{Q}_L Y^u U_R H_c, \quad (58)$$

( $H_c = i\tau_2 H^*$ ) break this symmetry. The Yukawa couplings can thus be thought of as spurions with the following transformation properties under  $SU(3)_q^3$  [see Eq. (15)]:

$$Y^u \sim (3, \bar{3}, 1), \quad Y^d \sim (3, 1, \bar{3}). \quad (59)$$

When we say ‘‘spurions’’, we mean that we pretend that the Yukawa matrices are fields which transform under the flavor symmetry, and then require that all the Lagrangian terms, constructed from the SM fields,  $Y^d$  and  $Y^u$ , must be (formally) invariant under the flavor group  $SU(3)_q^3$ . Of course, in reality,  $\mathcal{L}_Y$  breaks  $SU(3)_q^3$  precisely because  $Y^{d,u}$  are *not* fields and do not transform under the symmetry.

The idea of minimal flavor violation is relevant to extensions of the SM, and can be applied in two ways:

1. If we consider the SM as a low energy effective theory, then all higher-dimension operators, constructed from SM-fields and  $Y$ -spurions, are formally invariant under  $G_{\text{global}}$ .
2. If we consider a full high-energy theory that extends the SM, then all operators, constructed from SM and the new fields, and from  $Y$ -spurions, are formally invariant under  $G_{\text{global}}$ .

**Exercise 10:** Use the spurion formalism to argue that, in MFV models, the  $K_L \rightarrow \pi^0 \nu \bar{\nu}$  decay amplitude is proportional to  $y_t^2 V_{td} V_{ts}^*$ .

**Exercise 11:** Find the flavor suppression factors in the  $z_i^{bs}$  coefficients, if MFV is imposed, and compare to the bounds in Eq. (47).

Examples of MFV models include models of supersymmetry with gauge-mediation or with anomaly-mediation of its breaking.

### 5.3.1 Testing MFV at the LHC

If the LHC discovers new particles that couple to the SM fermions, then it will be able to test solutions to the new physics flavor puzzle such as MFV [32]. Much of its power to test such frameworks is based on identifying top and bottom quarks.

To understand this statement, we notice that the spurions  $Y^u$  and  $Y^d$  can always be written in terms of the two diagonal Yukawa matrices  $\lambda_u$  and  $\lambda_d$  and the CKM matrix  $V$ , see Eqs. (17,18). Thus, the only source of quark flavor changing transitions in MFV models is the CKM matrix. Next, note that to an accuracy that is better than  $\mathcal{O}(0.05)$ , we can write the CKM matrix as follows:

$$V = \begin{pmatrix} 1 & 0.23 & 0 \\ -0.23 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (60)$$

**Exercise 12:** *The approximation (60) should be intuitively obvious to top-physicists, but definitely counter-intuitive to bottom-physicists. (Some of them have dedicated a large part of their careers to experimental or theoretical efforts to determine  $V_{cb}$  and  $V_{ub}$ .) What does the approximation imply for the bottom quark? When we take into account that it is only good to  $\mathcal{O}(0.05)$ , what would the implications be?*

We learn that the third generation of quarks is decoupled, to a good approximation, from the first two. This, in turn, means that any new particle that couples to an odd number of the SM quarks (think, for example, of heavy quarks in vector-like representations of  $G_{\text{SM}}$ ), decay into either third generation quark, or to non-third generation quark, but not to both. For example, in Ref. [32], MFV models with additional charge  $-1/3$ ,  $SU(2)_L$ -singlet quarks  $-B'$  – were considered. A concrete test of MFV was proposed, based on the fact that the largest mixing effect involving the third generation is of order  $|V_{cb}|^2 \sim 0.002$ : Is the following prediction, concerning events of  $B'$  pair production, fulfilled:

$$\frac{\Gamma(B'\bar{B}' \rightarrow X q_{1,2} q_3)}{\Gamma(B'\bar{B}' \rightarrow X q_{1,2} q_{1,2}) + \Gamma(B'\bar{B}' \rightarrow X q_3 q_3)} \lesssim 10^{-3}. \quad (61)$$

If not, then MFV is excluded. One could similarly test various versions of minimal lepton flavor violation (MLFV) [33–38].

Analogous tests can be carried out in the supersymmetric framework [39–45]. Here, there is also a generic prediction that, in each of the three sectors ( $Q_L, U_R, D_R$ ), squarks of the first two generations are quasi-degenerate, and do not decay into third generation quarks. Squarks of the third generation can be separated in mass (though, for small  $\tan\beta$ , the degeneracy in the  $\tilde{D}_R$  sector is threefold), and decay only to third generation quarks.

We conclude that measurements at the LHC related to new particles that couple to the SM fermions are likely to teach us much more about flavor physics.

## 6 The Standard Model flavor puzzle

The SM has thirteen flavor parameters: six quark Yukawa couplings, four CKM parameters (three angles and a phase), and three charged lepton Yukawa couplings. (One can use fermion masses instead of the fermion Yukawa couplings,  $Y_f = \sqrt{2}m_f/v$ .) The orders of magnitudes of these thirteen dimensionless parameters are as follows:

$$\begin{aligned} Y_t &\sim 1, & Y_c &\sim 10^{-2}, & Y_u &\sim 10^{-5}, \\ Y_b &\sim 10^{-2}, & Y_s &\sim 10^{-3}, & Y_d &\sim 10^{-4}, \\ Y_\tau &\sim 10^{-2}, & Y_\mu &\sim 10^{-3}, & Y_e &\sim 10^{-6}, \\ |V_{us}| &\sim 0.2, & |V_{cb}| &\sim 0.04, & |V_{ub}| &\sim 0.004, & \delta_{\text{KM}} &\sim 1. \end{aligned} \quad (62)$$

Only two of these parameters are clearly of  $\mathcal{O}(1)$ , the top-Yukawa and the KM phase. The other flavor parameters exhibit smallness and hierarchy. Their values span six orders of magnitude. It may be that this set of numerical values are just accidental. More likely, the smallness and the hierarchy have a reason. The question of why there is smallness and hierarchy in the SM flavor parameters constitutes “The Standard Model flavor puzzle.”

The motivation to think that there is indeed a structure in the flavor parameters is strengthened by considering the values of the four SM parameters that are not flavor parameters, namely the three gauge couplings and the Higgs self-coupling:

$$g_s \sim 1, \quad g \sim 0.6, \quad e \sim 0.3, \quad \lambda \sim 0.2. \quad (63)$$

This set of values does seem to be a random distribution of order-one numbers, as one would naively expect.

A few examples of mechanisms that were proposed to explain the observed structure of the flavor parameters are the following:

- An approximate Abelian symmetry (“The Froggatt-Nielsen mechanism” [46]);
- An approximate non-Abelian symmetry (see *e.g.* [47]);
- Conformal dynamics (“The Nelson-Strassler mechanism” [48]);
- Location in an extra dimension [49].

We will take as an example the Froggatt-Nielsen mechanism.

### 6.1 The Froggatt-Nielsen mechanism

Small numbers and hierarchies are often explained by approximate symmetries. For example, the small mass splitting between the charged and neutral pions finds an explanation in the approximate isospin (global  $SU(2)$ ) symmetry of the strong interactions.

Approximate symmetries lead to selection rules which account for the size of deviations from the symmetry limit. Spurion analysis is particularly convenient to derive such selection rules. The Froggatt-Nielsen mechanism postulates a  $U(1)_H$  symmetry, that is broken by a small spurion  $\epsilon_H$ . Without loss of generality, we assign  $\epsilon_H$  a  $U(1)_H$  charge of  $H(\epsilon_H) = -1$ . Each SM field is assigned a  $U(1)_H$  charge. In general, different fermion generations are assigned different charges, hence the term ‘horizontal symmetry.’ The rule is that each term in the Lagrangian, made of SM fields and the spurion should be formally invariant under  $U(1)_H$ .

The approximate  $U(1)_H$  symmetry thus leads to the following selection rules:

$$\begin{aligned} Y_{ij}^u &= \epsilon_H^{|H(\bar{Q}_i)+H(U_j)+H(\phi_u)|}, \\ Y_{ij}^d &= \epsilon_H^{|H(\bar{Q}_i)+H(D_j)+H(\phi_d)|}, \\ Y_{ij}^e &= \epsilon_H^{|H(\bar{L}_i)+H(E_j)-H(\phi_d)|}. \end{aligned} \quad (64)$$

As a concrete example, we take the following set of charges:

$$\begin{aligned} H(\bar{Q}_i) &= H(U_i) = H(E_i) = (2, 1, 0), \\ H(\bar{L}_i) &= H(D_i) = (0, 0, 0), \\ H(\phi_u) &= H(\phi_d) = 0. \end{aligned} \quad (65)$$

It leads to the following parametric suppressions of the Yukawa couplings:

$$Y^u \sim \begin{pmatrix} \epsilon^4 & \epsilon^3 & \epsilon^2 \\ \epsilon^3 & \epsilon^2 & \epsilon \\ \epsilon^2 & \epsilon & 1 \end{pmatrix}, \quad Y^d \sim (Y^e)^T \sim \begin{pmatrix} \epsilon^2 & \epsilon^2 & \epsilon^2 \\ \epsilon & \epsilon & \epsilon \\ 1 & 1 & 1 \end{pmatrix}. \quad (66)$$

We emphasize that for each entry we give the parametric suppression (that is the power of  $\epsilon$ ), but each entry has an unknown (complex) coefficient of order one, and there are no relations between the order one coefficients of different entries.

The structure of the Yukawa matrices dictates the parametric suppression of the physical observables:

$$\begin{aligned} Y_t &\sim 1, & Y_c &\sim \epsilon^2, & Y_u &\sim \epsilon^4, \\ Y_b &\sim 1, & Y_s &\sim \epsilon, & Y_d &\sim \epsilon^2, \\ Y_\tau &\sim 1, & Y_\mu &\sim \epsilon, & Y_e &\sim \epsilon^2, \end{aligned}$$



$$|V_{us}| \sim \epsilon, \quad |V_{cb}| \sim \epsilon, \quad |V_{ub}| \sim \epsilon^2, \quad \delta_{\text{KM}} \sim 1. \quad (67)$$

For  $\epsilon \sim 0.05$ , the parametric suppressions are roughly consistent with the observed hierarchy. In particular, this set of charges predicts that the down and charged lepton mass hierarchies are similar, while the up hierarchy is the square of the down hierarchy. These features are roughly realized in Nature.

**Exercise 13:** *Derive the parametric suppression and approximate numerical values of  $Y^u$ , its eigenvalues, and the three angles of  $V_L^u$ , for  $H(Q_i) = 4, 2, 0$ ,  $H(U_i) = 3, 2, 0$  and  $\epsilon_H = 0.2$*

Could we explain any set of observed values with such an approximate symmetry? If we could, then the FN mechanism cannot be really tested. The answer however is negative. Consider, for example, the quark sector. Naively, we have 11  $U(1)_H$  charges that we are free to choose. However, the  $U(1)_Y \times U(1)_B \times U(1)_{\text{PQ}}$  symmetry implies that there are only 8 independent choices that affect the structure of the Yukawa couplings. On the other hand, there are 9 physical parameters. Thus, there should be a single relation between the physical parameters that is independent of the choice of charges. Assuming that the sum of charges in the exponents of Eq. (64) is of the same sign for all 18 combinations, the relation is

$$|V_{ub}| \sim |V_{us}V_{cb}|, \quad (68)$$

which is fulfilled to within a factor of 2. There are also interesting inequalities (here  $i < j$ ):

$$|V_{ij}| \gtrsim m(U_i)/m(U_j), \quad m(D_i)/m(D_j). \quad (69)$$

All six inequalities are fulfilled. Finally, if we order the up and the down masses from light to heavy, then the CKM matrix is predicted to be  $\sim \mathbf{1}$ , namely the diagonal entries are not parametrically suppressed. This structure is also consistent with the observed CKM structure.

## 6.2 The flavor of neutrinos

Five neutrino flavor parameters have been measured in recent years (see *e.g.* [50]): two mass-squared differences,

$$\Delta m_{21}^2 = (7.5 \pm 0.2) \times 10^{-5} \text{ eV}^2, \quad |\Delta m_{32}^2| = (2.5 \pm 0.1) \times 10^{-3} \text{ eV}^2, \quad (70)$$

and the three mixing angles,

$$|U_{e2}| = 0.55 \pm 0.01, \quad |U_{\mu 3}| = 0.64 \pm 0.02, \quad |U_{e3}| = 0.15 \pm 0.01. \quad (71)$$

These parameters constitute a significant addition to the thirteen SM flavor parameters and provide, in principle, tests of various ideas to explain the SM flavor puzzle.

The numerical values of the parameters show various surprising features:

- $|U_{\mu 3}| > \text{any } |V_{ij}|$ ;
- $|U_{e2}| > \text{any } |V_{ij}|$ ;
- $|U_{e3}|$  is not particularly small ( $|U_{e3}| \not\ll |U_{e2}U_{\mu 3}|$ );
- $m_2/m_3 \gtrsim 1/6 > \text{any } m_i/m_j$  for charged fermions.

These features can be summarized by the statement that, in contrast to the charged fermions, neither smallness nor hierarchy have been observed so far in the neutrino related parameters.

One way of interpretation of the neutrino data comes under the name of neutrino mass anarchy [51–53]. It postulates that the neutrino mass matrix has no structure, namely all entries are of the same order of magnitude. Normalized to an effective neutrino mass scale,  $v^2/\Lambda_{\text{seesaw}}$ , the various entries are random numbers of order one. Note that anarchy means neither hierarchy nor degeneracy.

If true, the contrast between neutrino mass anarchy and quark and charged lepton mass hierarchy may be a deep hint for a difference between the flavor physics of Majorana and Dirac fermions. The source of both anarchy and hierarchy might, however, be explained by a much more mundane mechanism. In particular, neutrino mass anarchy could be a result of a FN mechanism, where the three left-handed lepton doublets carry the same FN charge. In that case, the FN mechanism predict parametric suppression of neither neutrino mass ratios nor leptonic mixing angles, which is quite consistent with (70) and (71). Indeed, the viable FN model presented in Section 6.1 belongs to this class.

Another possible interpretation of the neutrino data is to take  $m_2/m_3 \sim |U_{e3}| \sim 0.15$  to be small, and require that they are parametrically suppressed (while the other two mixing angles are order one). Such a situation is impossible to accommodate in a large class of FN models [54].

The same data, and in particular the proximity of  $|U_{e2}|$  to  $1/\sqrt{3} \simeq 0.58$  and the proximity of  $|U_{\mu 3}|$  to  $1/\sqrt{2} \simeq 0.71$  led to a very different interpretation. This interpretation, termed ‘tribimaximal mixing’ (TBM), postulates that the leptonic mixing matrix is parametrically close to the following special form [55]:

$$|U|_{\text{TBM}} = \begin{pmatrix} \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \end{pmatrix}. \quad (72)$$

Such a form is suggestive of discrete non-Abelian symmetries, and indeed numerous models based on an  $A_4$  symmetry have been proposed [56, 57]. A significant feature of TBM is that the third mixing angle should be close to  $|U_{e3}| = 0$ . Until recently, there have been only upper bounds on  $|U_{e3}|$ , consistent with the models in the literature. In the last year, however, a value of  $|U_{e3}|$  close to the previous upper bound has been established [58], see Eq. (71). Such a large value (and the consequent significant deviation of  $|U_{\mu 3}|$  from maximal bimixing) puts in serious doubt the TBM idea. Indeed, it is difficult in this framework, if not impossible, to account for  $\Delta m_{12}^2/\Delta m_{23}^2 \sim |U_{e3}|^2$  without fine-tuning [59].

## 7 Higgs physics: the new flavor arena

A Higgs-like boson  $h$  has been discovered by the ATLAS and CMS experiments at the LHC [60, 61]. The fact that for the  $f = \gamma\gamma$  and  $f = ZZ^*$  final states, the experiments measure

$$R_f \equiv \frac{\sigma(pp \rightarrow h)\text{BR}(h \rightarrow f)}{[\sigma(pp \rightarrow h)\text{BR}(h \rightarrow f)]^{\text{SM}}}, \quad (73)$$

of order one (see *e.g.* [62]),

$$R_{ZZ^*} = 1.1 \pm 0.2, \quad (74)$$

$$R_{\gamma\gamma} = 1.1 \pm 0.2, \quad (75)$$

is suggestive that the  $h$ -production via gluon-gluon fusion proceeds at a rate similar to the Standard Model (SM) prediction, giving a strong indication that  $Y_t$ , the  $h t \bar{t}$  Yukawa coupling, is of order one. This first determination of  $Y_t$  signifies a new arena for the exploration of *flavor physics*.

In the future, measurements of  $R_{b\bar{b}}$  and  $R_{\tau^+\tau^-}$  will allow us to extract additional flavor parameters:  $Y_b$ , the  $h b \bar{b}$  Yukawa coupling, and  $Y_\tau$ , the  $h \tau^+ \tau^-$  Yukawa coupling. For the latter, the current allowed range is already quite restrictive:

$$R_{\tau^+\tau^-} = 1.0 \pm 0.4. \quad (76)$$

It may well be that the values of  $Y_b$  and/or  $Y_\tau$  will deviate from their SM values. The most likely explanation of such deviations will be that there are more than one Higgs doublets, and that the doublet(s) that couple to the down and charged lepton sectors are not the same as the one that couples to the up sector.

A more significant test of our understanding of flavor physics, which might provide a window into new flavor physics, will come further in the future, when  $R_{\mu^+\mu^-}$  is measured. (At present, there is an upper bound,  $R_{\mu^+\mu^-} < 9.8$ .) The ratio

$$X_{\mu^+\mu^-} \equiv \frac{\text{BR}(h \rightarrow \mu^+\mu^-)}{\text{BR}(h \rightarrow \tau^+\tau^-)}, \quad (77)$$

is predicted within the SM with impressive theoretical cleanliness. To leading order, it is given by  $X_{\mu^+\mu^-} = m_\mu^2/m_\tau^2$ , and the corrections of order  $\alpha_W$  and of order  $m_\mu^2/m_\tau^2$  to this leading result are known. It is an interesting question to understand what can be learned from a test of this relation [63,64].

It is also possible to search for the SM-forbidden decay modes,  $h \rightarrow \mu^\pm\tau^\mp$  [65–68]. A measurement of, or an upper bound on

$$X_{\mu\tau} \equiv \frac{\text{BR}(h \rightarrow \mu^+\tau^-) + \text{BR}(h \rightarrow \mu^-\tau^+)}{\text{BR}(h \rightarrow \tau^+\tau^-)}, \quad (78)$$

would provide additional information relevant to flavor physics. Thus, a broader question is to understand the implications for flavor physics of measurements of  $R_{\tau^+\tau^-}$ ,  $X_{\mu^+\mu^-}$  and  $X_{\mu\tau}$  [63].

Let us take as an example how we can use the set of these three measurements if there is a single light Higgs boson. A violation of the SM relation  $Y_{ij}^{\text{SM}} = \frac{\sqrt{2}m_i}{v}\delta_{ij}$ , is a consequence of nonrenormalizable terms. The leading ones are the  $d = 6$  terms. In the interaction basis, we have

$$\begin{aligned} \mathcal{L}_Y^{d=4} &= -\lambda_{ij}\bar{f}_L^i f_R^j \phi + \text{h.c.}, \\ \mathcal{L}_Y^{d=6} &= -\frac{\lambda'_{ij}}{\Lambda^2}\bar{f}_L^i f_R^j \phi(\phi^\dagger\phi) + \text{h.c.}, \end{aligned} \quad (79)$$

where expanding around the vacuum we have  $\phi = (v + h)/\sqrt{2}$ . Defining  $V_{L,R}$  via

$$\sqrt{2}m = V_L \left( \lambda + \frac{v^2}{2\Lambda^2}\lambda' \right) V_R^\dagger v, \quad (80)$$

where  $m = \text{diag}(m_e, m_\mu, m_\tau)$ , and defining  $\hat{\lambda}$  via

$$\hat{\lambda} = V_L \lambda' V_R^\dagger, \quad (81)$$

we obtain

$$Y_{ij} = \frac{\sqrt{2}m_i}{v}\delta_{ij} + \frac{v^2}{\Lambda^2}\hat{\lambda}_{ij}. \quad (82)$$

To proceed, one has to make assumptions about the structure of  $\hat{\lambda}$ . In what follows, we consider first the assumption of minimal flavor violation (MFV) and then a Froggatt-Nielsen (FN) symmetry.

## 7.1 MFV

MFV requires that the leptonic part of the Lagrangian is invariant under an  $SU(3)_L \times SU(3)_E$  global symmetry, with the left-handed lepton doublets transforming as  $(3, 1)$ , the right-handed charged lepton singlets transforming as  $(1, 3)$  and the charged lepton Yukawa matrix  $Y$  is a spurion transforming as  $(3, \bar{3})$ .

Specifically, MFV means that, in Eq. (79),

$$\lambda' = a\lambda + b\lambda\lambda^\dagger\lambda + \mathcal{O}(\lambda^5), \quad (83)$$

where  $a$  and  $b$  are numbers. Note that, if  $V_L$  and  $V_R$  are the diagonalizing matrices for  $\lambda$ ,  $V_L \lambda V_R^\dagger = \lambda^{\text{diag}}$ , then they are also the diagonalizing matrices for  $\lambda \lambda^\dagger \lambda$ ,  $V_L \lambda \lambda^\dagger \lambda V_R^\dagger = (\lambda^{\text{diag}})^3$ . Then, Eqs. (80), (81) and (82) become

$$\begin{aligned} \frac{\sqrt{2}m}{v} &= \left(1 + \frac{av^2}{2\Lambda^2}\right) \lambda^{\text{diag}} + \frac{bv^2}{2\Lambda^2} (\lambda^{\text{diag}})^3, \\ \hat{\lambda} &= a\lambda^{\text{diag}} + b(\lambda^{\text{diag}})^3 = a \frac{\sqrt{2}m}{v} + \frac{2\sqrt{2}bm^3}{v^3}, \\ Y_{ij} &= \frac{\sqrt{2}m_i}{v} \delta_{ij} \left[1 + \frac{av^2}{\Lambda^2} + \frac{2bm_i^2}{\Lambda^2}\right], \end{aligned} \quad (84)$$

where, in the expressions for  $\hat{\lambda}$  and  $Y$ , we included only the leading universal and leading non-universal corrections to the SM relations.

We learn the following points about the Higgs-related lepton flavor parameters in this class of models:

1.  $h$  has no flavor off-diagonal couplings:

$$Y_{\mu\tau}, Y_{\tau\mu} = 0. \quad (85)$$

2. The values of the diagonal couplings deviate from their SM values. The deviation is small, of order  $v^2/\Lambda^2$ :

$$Y_\tau \approx \left(1 + \frac{av^2}{\Lambda^2}\right) \frac{\sqrt{2}m_\tau}{v}. \quad (86)$$

3. The ratio between the Yukawa couplings to different charged lepton flavors deviates from its SM value. The deviation is, however, very small, of order  $m_\tau^2/\Lambda^2$ :

$$\frac{Y_\mu}{Y_\tau} = \frac{m_\mu}{m_\tau} \left(1 - \frac{2b(m_\tau^2 - m_\mu^2)}{\Lambda^2}\right). \quad (87)$$

The predictions of the SM with MFV non-renormalizable terms are then the following:

$$\begin{aligned} \left(\frac{\sigma(pp \rightarrow h)^{\text{SM}}}{\sigma(pp \rightarrow h)} \frac{\Gamma_{\text{tot}}}{\Gamma_{\text{tot}}^{\text{SM}}}\right) R_{\tau^+\tau^-} &= 1 + 2av^2/\Lambda^2, \\ X_{\mu^+\mu^-} &= (m_\mu/m_\tau)^2(1 - 4bm_\tau^2/\Lambda^2), \\ X_{\tau\mu} &= 0. \end{aligned} \quad (88)$$

Thus, MFV will be excluded if experiments observe the  $h \rightarrow \mu\tau$  decay. On the other hand, MFV allows for a universal deviation of  $\mathcal{O}(v^2/\Lambda^2)$  of the flavor-diagonal dilepton rates, and a smaller non-universal deviation of  $\mathcal{O}(m_\tau^2/\Lambda^2)$ .

## 7.2 FN

An attractive explanation of the smallness and hierarchy in the Yukawa couplings is provided by the Froggatt-Nielsen (FN) mechanism [46]. In this framework, a  $U(1)_H$  symmetry, under which different generations carry different charges, is broken by a small parameter  $\epsilon_H$ . Without loss of generality,  $\epsilon_H$  is taken to be a spurion of charge  $-1$ . Then, various entries in the Yukawa mass matrices are suppressed by different powers of  $\epsilon_H$ , leading to smallness and hierarchy.

Specifically for the leptonic Yukawa matrix, taking  $h$  to be neutral under  $U(1)_H$ ,  $H(h) = 0$ , we have

$$\lambda_{ij} \propto \epsilon_H^{H(E_j) - H(L_i)}. \quad (89)$$

We emphasize that the FN mechanism dictates only the parametric suppression. Each entry has an arbitrary order one coefficient. The resulting parametric suppression of the masses and leptonic mixing angles is given by [69]

$$m_{\ell_i}/v \sim \epsilon_H^{H(E_i)-H(L_i)}, \quad |U_{ij}| \sim \epsilon_H^{H(L_j)-H(L_i)}. \quad (90)$$

Since  $H(\phi^\dagger\phi) = 0$ , the entries of the matrix  $\lambda'$  have the same parametric suppression as the corresponding entries in  $\lambda$  [26], though the order one coefficients are different:

$$\lambda'_{ij} = \mathcal{O}(1) \times \lambda_{ij}. \quad (91)$$

This structure allows us to estimate the entries of  $\hat{\lambda}_{ij}$  in terms of physical observables:

$$\begin{aligned} \hat{\lambda}_{33} &\sim m_\tau/v, \\ \hat{\lambda}_{22} &\sim m_\mu/v, \\ \hat{\lambda}_{23} &\sim |U_{23}|(m_\tau/v), \\ \hat{\lambda}_{32} &\sim (m_\mu/v)/|U_{23}|. \end{aligned} \quad (92)$$

We learn the following points about the Higgs-related lepton flavor parameters in this class of models:

1.  $h$  has flavor off-diagonal couplings:

$$\begin{aligned} Y_{\mu\tau} &= \mathcal{O}\left(\frac{|U_{23}|vm_\tau}{\Lambda^2}\right), \\ Y_{\tau\mu} &= \mathcal{O}\left(\frac{vm_\mu}{|U_{23}|\Lambda^2}\right). \end{aligned} \quad (93)$$

2. The values of the diagonal couplings deviate from their SM values:

$$Y_\tau \approx \frac{\sqrt{2}m_\tau}{v} \left[ 1 + \mathcal{O}\left(\frac{v^2}{\Lambda^2}\right) \right]. \quad (94)$$

3. The ratio between the Yukawa couplings to different charged lepton flavors deviates from its SM value:

$$\frac{Y_\mu}{Y_\tau} = \frac{m_\mu}{m_\tau} \left[ 1 + \mathcal{O}\left(\frac{v^2}{\Lambda^2}\right) \right]. \quad (95)$$

The predictions of the SM with FN-suppressed non-renormalizable terms are then the following:

$$\begin{aligned} \left( \frac{\sigma(pp \rightarrow h)^{\text{SM}}}{\sigma(pp \rightarrow h)} \frac{\Gamma_{\text{tot}}}{\Gamma_{\text{tot}}^{\text{SM}}} \right) R_{\tau^+\tau^-} &= 1 + \mathcal{O}(v^2/\Lambda^2), \\ X_{\mu^+\mu^-} &= (m_\mu/m_\tau)^2(1 + \mathcal{O}(v^2/\Lambda^2)), \\ X_{\tau\mu} &= \mathcal{O}(v^4/\Lambda^4). \end{aligned} \quad (96)$$

Thus, FN will be excluded if experiments observe deviations from the SM of the same size in both flavor-diagonal and flavor-changing  $h$  decays. On the other hand, FN allows non-universal deviations of  $\mathcal{O}(v^2/\Lambda^2)$  in the flavor-diagonal dilepton rates, and a smaller deviation of  $\mathcal{O}(v^4/\Lambda^4)$  in the off-diagonal rate.

## 8 Conclusions

(i) Measurements of CP violating  $B$ -meson decays have established that the Kobayashi-Maskawa mechanism is the dominant source of the observed CP violation.

(ii) Measurements of flavor changing  $B$ -meson decays have established the the Cabibbo-Kobayashi-Maskawa mechanism is a major player in flavor violation.

(iii) The consistency of all these measurements with the CKM predictions sharpens the new physics flavor puzzle: If there is new physics at, or below, the TeV scale, then its flavor structure must be highly non-generic.

(iv) Measurements of neutrino flavor parameters have not only not clarified the standard model flavor puzzle, but actually deepened it. Whether they imply an anarchical structure, or a tribimaximal mixing, it seems that the neutrino flavor structure is very different from that of quarks.

(v) If the LHC experiments, ATLAS and CMS, discover new particles that couple to the Standard Model fermions, then, in principle, they will be able to measure new flavor parameters. Consequently, the new physics flavor puzzle is likely to be understood.

(vi) If the flavor structure of such new particles is affected by the same physics that sets the flavor structure of the Yukawa couplings, then the LHC experiments (and future flavor factories) may be able to shed light also on the standard model flavor puzzle.

(vii) The recently discovered Higgs-like boson provides an opportunity to make progress in our understanding of the flavor puzzle(s).

The huge progress in flavor physics in recent years has provided answers to many questions. At the same time, new questions arise. The LHC era is likely to provide more answers and more questions.

## Appendices

### A The CKM matrix

The CKM matrix  $V$  is a  $3 \times 3$  unitary matrix. Its form, however, is not unique:

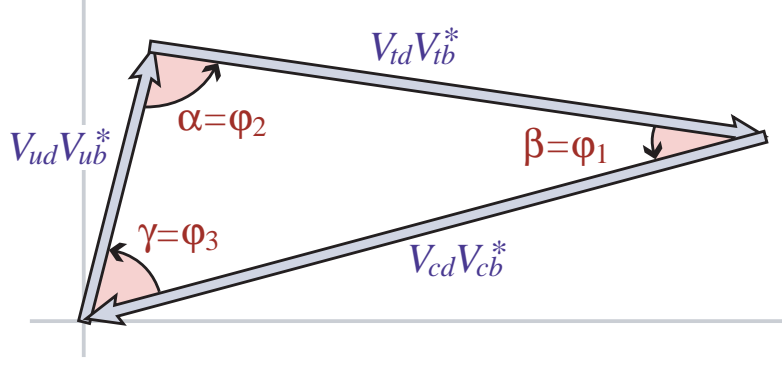
(i) There is freedom in defining  $V$  in that we can permute between the various generations. This freedom is fixed by ordering the up quarks and the down quarks by their masses, *i.e.*  $(u_1, u_2, u_3) \rightarrow (u, c, t)$  and  $(d_1, d_2, d_3) \rightarrow (d, s, b)$ . The elements of  $V$  are written as follows:

$$V = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}. \quad (\text{A.1})$$

(ii) There is further freedom in the phase structure of  $V$ . This means that the number of physical parameters in  $V$  is smaller than the number of parameters in a general unitary  $3 \times 3$  matrix which is nine (three real angles and six phases). Let us define  $P_q$  ( $q = u, d$ ) to be diagonal unitary (phase) matrices. Then, if instead of using  $V_{qL}$  and  $V_{qR}$  for the rotation (21) to the mass basis we use  $\tilde{V}_{qL}$  and  $\tilde{V}_{qR}$ , defined by  $\tilde{V}_{qL} = P_q V_{qL}$  and  $\tilde{V}_{qR} = P_q V_{qR}$ , we still maintain a legitimate mass basis since  $M_q^{\text{diag}}$  remains unchanged by such transformations. However,  $V$  does change:

$$V \rightarrow P_u V P_d^*. \quad (\text{A.2})$$

This freedom is fixed by demanding that  $V$  has the minimal number of phases. In the three generation case  $V$  has a single phase. (There are five phase differences between the elements of  $P_u$  and  $P_d$  and, therefore, five of the six phases in the CKM matrix can be removed.) This is the Kobayashi-Maskawa phase  $\delta_{\text{KM}}$  which is the single source of CP violation in the quark sector of the Standard Model [1].



**Fig. A.1:** Graphical representation of the unitarity constraint  $V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0$  as a triangle in the complex plane.

The fact that  $V$  is unitary and depends on only four independent physical parameters can be made manifest by choosing a specific parametrization. The standard choice is [70]

$$V = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix}, \quad (\text{A.3})$$

where  $c_{ij} \equiv \cos \theta_{ij}$  and  $s_{ij} \equiv \sin \theta_{ij}$ . The  $\theta_{ij}$ 's are the three real mixing parameters while  $\delta$  is the Kobayashi-Maskawa phase. It is known experimentally that  $s_{13} \ll s_{23} \ll s_{12} \ll 1$ . It is convenient to choose an approximate expression where this hierarchy is manifest. This is the Wolfenstein parametrization, where the four mixing parameters are  $(\lambda, A, \rho, \eta)$  with  $\lambda = |V_{us}| = 0.23$  playing the role of an expansion parameter and  $\eta$  representing the CP violating phase [71, 72]:

$$V = \begin{pmatrix} 1 - \frac{1}{2}\lambda^2 - \frac{1}{8}\lambda^4 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda + \frac{1}{2}A^2\lambda^5[1 - 2(\rho + i\eta)] & 1 - \frac{1}{2}\lambda^2 - \frac{1}{8}\lambda^4(1 + 4A^2) & A\lambda^2 \\ A\lambda^3[1 - (1 - \frac{1}{2}\lambda^2)(\rho + i\eta)] & -A\lambda^2 + \frac{1}{2}A\lambda^4[1 - 2(\rho + i\eta)] & 1 - \frac{1}{2}A^2\lambda^4 \end{pmatrix}. \quad (\text{A.4})$$

A very useful concept is that of the *unitarity triangles*. The unitarity of the CKM matrix leads to various relations among the matrix elements, *e.g.*

$$V_{ud}V_{us}^* + V_{cd}V_{cs}^* + V_{td}V_{ts}^* = 0, \quad (\text{A.5})$$

$$V_{us}V_{ub}^* + V_{cs}V_{cb}^* + V_{ts}V_{tb}^* = 0, \quad (\text{A.6})$$

$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0. \quad (\text{A.7})$$

Each of these three relations requires the sum of three complex quantities to vanish and so can be geometrically represented in the complex plane as a triangle. These are “the unitarity triangles”, though the term “unitarity triangle” is usually reserved for the relation (A.7) only. The unitarity triangle related to Eq. (A.7) is depicted in Fig. A.1.

The rescaled unitarity triangle is derived from (A.7) by (a) choosing a phase convention such that  $(V_{cd}V_{cb}^*)$  is real, and (b) dividing the lengths of all sides by  $|V_{cd}V_{cb}^*|$ . Step (a) aligns one side of the triangle with the real axis, and step (b) makes the length of this side 1. The form of the triangle is unchanged. Two vertices of the rescaled unitarity triangle are thus fixed at  $(0,0)$  and  $(1,0)$ . The coordinates of the remaining vertex correspond to the Wolfenstein parameters  $(\rho, \eta)$ . The area of the rescaled unitarity triangle is  $|\eta|/2$ .

Depicting the rescaled unitarity triangle in the  $(\rho, \eta)$  plane, the lengths of the two complex sides are

$$R_u \equiv \left| \frac{V_{ud}V_{ub}^*}{V_{cd}V_{cb}^*} \right| = \sqrt{\rho^2 + \eta^2}, \quad R_t \equiv \left| \frac{V_{td}V_{tb}^*}{V_{cd}V_{cb}^*} \right| = \sqrt{(1 - \rho)^2 + \eta^2}. \quad (\text{A.8})$$

The three angles of the unitarity triangle are defined as follows [73, 74]:

$$\alpha \equiv \arg \left[ -\frac{V_{td}V_{tb}^*}{V_{ud}V_{ub}^*} \right], \quad \beta \equiv \arg \left[ -\frac{V_{cd}V_{cb}^*}{V_{td}V_{tb}^*} \right], \quad \gamma \equiv \arg \left[ -\frac{V_{ud}V_{ub}^*}{V_{cd}V_{cb}^*} \right]. \quad (\text{A.9})$$

They are physical quantities and can be independently measured by CP asymmetries in  $B$  decays. It is also useful to define the two small angles of the unitarity triangles (A.6,A.5):

$$\beta_s \equiv \arg \left[ -\frac{V_{ts}V_{tb}^*}{V_{cs}V_{cb}^*} \right], \quad \beta_K \equiv \arg \left[ -\frac{V_{cs}V_{cd}^*}{V_{us}V_{ud}^*} \right]. \quad (\text{A.10})$$

## B CPV in $B$ decays to final CP eigenstates

We define decay amplitudes of  $B$  (which could be charged or neutral) and its CP conjugate  $\bar{B}$  to a multi-particle final state  $f$  and its CP conjugate  $\bar{f}$  as

$$A_f = \langle f | \mathcal{H} | B \rangle, \quad \bar{A}_f = \langle f | \mathcal{H} | \bar{B} \rangle, \quad A_{\bar{f}} = \langle \bar{f} | \mathcal{H} | B \rangle, \quad \bar{A}_{\bar{f}} = \langle \bar{f} | \mathcal{H} | \bar{B} \rangle, \quad (\text{B.1})$$

where  $\mathcal{H}$  is the Hamiltonian governing weak interactions. The action of CP on these states introduces phases  $\xi_B$  and  $\xi_f$  according to

$$\begin{aligned} CP |B\rangle &= e^{+i\xi_B} |\bar{B}\rangle, & CP |f\rangle &= e^{+i\xi_f} |\bar{f}\rangle, \\ CP |\bar{B}\rangle &= e^{-i\xi_B} |B\rangle, & CP |\bar{f}\rangle &= e^{-i\xi_f} |f\rangle, \end{aligned} \quad (\text{B.2})$$

so that  $(CP)^2 = 1$ . The phases  $\xi_B$  and  $\xi_f$  are arbitrary and unphysical because of the flavor symmetry of the strong interaction. If CP is conserved by the dynamics,  $[CP, \mathcal{H}] = 0$ , then  $A_f$  and  $\bar{A}_{\bar{f}}$  have the same magnitude and an arbitrary unphysical relative phase

$$\bar{A}_{\bar{f}} = e^{i(\xi_f - \xi_B)} A_f. \quad (\text{B.3})$$

A state that is initially a superposition of  $B^0$  and  $\bar{B}^0$ , say

$$|\psi(0)\rangle = a(0)|B^0\rangle + b(0)|\bar{B}^0\rangle, \quad (\text{B.4})$$

will evolve in time acquiring components that describe all possible decay final states  $\{f_1, f_2, \dots\}$ , that is,

$$|\psi(t)\rangle = a(t)|B^0\rangle + b(t)|\bar{B}^0\rangle + c_1(t)|f_1\rangle + c_2(t)|f_2\rangle + \dots. \quad (\text{B.5})$$

If we are interested in computing only the values of  $a(t)$  and  $b(t)$  (and not the values of all  $c_i(t)$ ), and if the times  $t$  in which we are interested are much larger than the typical strong interaction scale, then we can use a much simplified formalism [75]. The simplified time evolution is determined by a  $2 \times 2$  effective Hamiltonian  $\mathcal{H}$  that is not Hermitian, since otherwise the mesons would only oscillate and not decay. Any complex matrix, such as  $\mathcal{H}$ , can be written in terms of Hermitian matrices  $M$  and  $\Gamma$  as

$$\mathcal{H} = M - \frac{i}{2} \Gamma. \quad (\text{B.6})$$

$M$  and  $\Gamma$  are associated with  $(B^0, \bar{B}^0) \leftrightarrow (B^0, \bar{B}^0)$  transitions via off-shell (dispersive) and on-shell (absorptive) intermediate states, respectively. Diagonal elements of  $M$  and  $\Gamma$  are associated with the flavor-conserving transitions  $B^0 \rightarrow B^0$  and  $\bar{B}^0 \rightarrow \bar{B}^0$  while off-diagonal elements are associated with flavor-changing transitions  $B^0 \leftrightarrow \bar{B}^0$ .

The eigenvectors of  $\mathcal{H}$  have well defined masses and decay widths. We introduce complex parameters  $p$  and  $q$  to specify the components of the strong interaction eigenstates,  $B^0$  and  $\bar{B}^0$ , in the light ( $B_L$ ) and heavy ( $B_H$ ) mass eigenstates:

$$|B_{L,H}\rangle = p|B^0\rangle \pm q|\bar{B}^0\rangle \quad (\text{B.7})$$



with the normalization  $|p|^2 + |q|^2 = 1$ . The special form of Eq. (B.7) is related to the fact that CPT imposes  $M_{11} = M_{22}$  and  $\Gamma_{11} = \Gamma_{22}$ . Solving the eigenvalue problem gives

$$\left(\frac{q}{p}\right)^2 = \frac{M_{12}^* - (i/2)\Gamma_{12}^*}{M_{12} - (i/2)\Gamma_{12}}. \quad (\text{B.8})$$

If either CP or T is a symmetry of  $\mathcal{H}$ , then  $M_{12}$  and  $\Gamma_{12}$  are relatively real, leading to

$$\left(\frac{q}{p}\right)^2 = e^{2i\xi_B} \Rightarrow \left|\frac{q}{p}\right| = 1, \quad (\text{B.9})$$

where  $\xi_B$  is the arbitrary unphysical phase introduced in Eq. (B.2).

The real and imaginary parts of the eigenvalues of  $\mathcal{H}$  corresponding to  $|B_{L,H}\rangle$  represent their masses and decay-widths, respectively. The mass difference  $\Delta m_B$  and the width difference  $\Delta\Gamma_B$  are defined as follows:

$$\Delta m_B \equiv M_H - M_L, \quad \Delta\Gamma_B \equiv \Gamma_H - \Gamma_L. \quad (\text{B.10})$$

Note that here  $\Delta m_B$  is positive by definition, while the sign of  $\Delta\Gamma_B$  is to be experimentally determined. The average mass and width are given by

$$m_B \equiv \frac{M_H + M_L}{2}, \quad \Gamma_B \equiv \frac{\Gamma_H + \Gamma_L}{2}. \quad (\text{B.11})$$

It is useful to define dimensionless ratios  $x$  and  $y$ :

$$x \equiv \frac{\Delta m_B}{\Gamma_B}, \quad y \equiv \frac{\Delta\Gamma_B}{2\Gamma_B}. \quad (\text{B.12})$$

Solving the eigenvalue equation gives

$$(\Delta m_B)^2 - \frac{1}{4}(\Delta\Gamma_B)^2 = (4|M_{12}|^2 - |\Gamma_{12}|^2), \quad \Delta m_B \Delta\Gamma_B = 4\mathcal{R}e(M_{12}\Gamma_{12}^*). \quad (\text{B.13})$$

All CP-violating observables in  $B$  and  $\bar{B}$  decays to final states  $f$  and  $\bar{f}$  can be expressed in terms of phase-convention-independent combinations of  $A_f, \bar{A}_f, A_{\bar{f}}$  and  $\bar{A}_{\bar{f}}$ , together with, for neutral-meson decays only,  $q/p$ . CP violation in charged-meson decays depends only on the combination  $|\bar{A}_{\bar{f}}/A_f|$ , while CP violation in neutral-meson decays is complicated by  $B^0 \leftrightarrow \bar{B}^0$  oscillations and depends, additionally, on  $|q/p|$  and on  $\lambda_f \equiv (q/p)(\bar{A}_f/A_f)$ .

For neutral  $D$ ,  $B$ , and  $B_s$  mesons,  $\Delta\Gamma/\Gamma \ll 1$  and so both mass eigenstates must be considered in their evolution. We denote the state of an initially pure  $|B^0\rangle$  or  $|\bar{B}^0\rangle$  after an elapsed proper time  $t$  as  $|B^0_{\text{phys}}(t)\rangle$  or  $|\bar{B}^0_{\text{phys}}(t)\rangle$ , respectively. Using the effective Hamiltonian approximation, we obtain

$$\begin{aligned} |B^0_{\text{phys}}(t)\rangle &= g_+(t)|B^0\rangle - \frac{q}{p}g_-(t)|\bar{B}^0\rangle, \\ |\bar{B}^0_{\text{phys}}(t)\rangle &= g_+(t)|\bar{B}^0\rangle - \frac{p}{q}g_-(t)|B^0\rangle, \end{aligned} \quad (\text{B.14})$$

where

$$g_{\pm}(t) \equiv \frac{1}{2} \left( e^{-im_H t - \frac{1}{2}\Gamma_H t} \pm e^{-im_L t - \frac{1}{2}\Gamma_L t} \right). \quad (\text{B.15})$$

One obtains the following time-dependent decay rates:

$$\frac{d\Gamma[B^0_{\text{phys}}(t) \rightarrow f]/dt}{e^{-\Gamma t}\mathcal{N}_f} = (|A_f|^2 + |(q/p)\bar{A}_f|^2) \cosh(y\Gamma t) + (|A_f|^2 - |(q/p)\bar{A}_f|^2) \cos(x\Gamma t)$$

$$+ 2 \operatorname{Re}((q/p)A_f^* \bar{A}_f) \sinh(y\Gamma t) - 2 \operatorname{Im}((q/p)A_f^* \bar{A}_f) \sin(x\Gamma t), \quad (\text{B.16})$$

$$\begin{aligned} \frac{d\Gamma[\bar{B}_{\text{phys}}^0(t) \rightarrow f]/dt}{e^{-\Gamma t} \mathcal{N}_f} &= (|(p/q)A_f|^2 + |\bar{A}_f|^2) \cosh(y\Gamma t) - (|(p/q)A_f|^2 - |\bar{A}_f|^2) \cos(x\Gamma t) \\ &+ 2 \operatorname{Re}((p/q)A_f \bar{A}_f^*) \sinh(y\Gamma t) - 2 \operatorname{Im}((p/q)A_f \bar{A}_f^*) \sin(x\Gamma t), \quad (\text{B.17}) \end{aligned}$$

where  $\mathcal{N}_f$  is a common normalization factor. Decay rates to the CP-conjugate final state  $\bar{f}$  are obtained analogously, with  $\mathcal{N}_f = \mathcal{N}_{\bar{f}}$  and the substitutions  $A_f \rightarrow A_{\bar{f}}$  and  $\bar{A}_f \rightarrow \bar{A}_{\bar{f}}$  in Eqs. (B.16,B.17). Terms proportional to  $|A_f|^2$  or  $|\bar{A}_f|^2$  are associated with decays that occur without any net  $B \leftrightarrow \bar{B}$  oscillation, while terms proportional to  $|(q/p)\bar{A}_f|^2$  or  $|(p/q)A_f|^2$  are associated with decays following a net oscillation. The  $\sinh(y\Gamma t)$  and  $\sin(x\Gamma t)$  terms of Eqs. (B.16,B.17) are associated with the interference between these two cases. Note that, in multi-body decays, amplitudes are functions of phase-space variables. Interference may be present in some regions but not others, and is strongly influenced by resonant substructure.

One possible manifestation of CP-violating effects in meson decays [76] is in the interference between a decay without mixing,  $B^0 \rightarrow f$ , and a decay with mixing,  $B^0 \rightarrow \bar{B}^0 \rightarrow f$  (such an effect occurs only in decays to final states that are common to  $B^0$  and  $\bar{B}^0$ , including all CP eigenstates). It is defined by

$$\operatorname{Im}(\lambda_f) \neq 0, \quad (\text{B.18})$$

with

$$\lambda_f \equiv \frac{q \bar{A}_f}{p A_f}. \quad (\text{B.19})$$

This form of CP violation can be observed, for example, using the asymmetry of neutral meson decays into final CP eigenstates  $f_{CP}$

$$\mathcal{A}_{f_{CP}}(t) \equiv \frac{d\Gamma/dt[\bar{B}_{\text{phys}}^0(t) \rightarrow f_{CP}] - d\Gamma/dt[B_{\text{phys}}^0(t) \rightarrow f_{CP}]}{d\Gamma/dt[\bar{B}_{\text{phys}}^0(t) \rightarrow f_{CP}] + d\Gamma/dt[B_{\text{phys}}^0(t) \rightarrow f_{CP}]}. \quad (\text{B.20})$$

For  $\Delta\Gamma = 0$  and  $|q/p| = 1$  (which is a good approximation for  $B$  mesons),  $\mathcal{A}_{f_{CP}}$  has a particularly simple form [77–79]:

$$\begin{aligned} \mathcal{A}_f(t) &= S_f \sin(\Delta m t) - C_f \cos(\Delta m t), \\ S_f &\equiv \frac{2 \operatorname{Im}(\lambda_f)}{1 + |\lambda_f|^2}, \quad C_f \equiv \frac{1 - |\lambda_f|^2}{1 + |\lambda_f|^2}, \quad (\text{B.21}) \end{aligned}$$

Consider the  $B \rightarrow f$  decay amplitude  $A_f$ , and the CP conjugate process,  $\bar{B} \rightarrow \bar{f}$ , with decay amplitude  $\bar{A}_{\bar{f}}$ . There are two types of phases that may appear in these decay amplitudes. Complex parameters in any Lagrangian term that contributes to the amplitude will appear in complex conjugate form in the CP-conjugate amplitude. Thus their phases appear in  $A_f$  and  $\bar{A}_{\bar{f}}$  with opposite signs. In the Standard Model, these phases occur only in the couplings of the  $W^\pm$  bosons and hence are often called “weak phases”. The weak phase of any single term is convention dependent. However, the difference between the weak phases in two different terms in  $A_f$  is convention independent. A second type of phase can appear in scattering or decay amplitudes even when the Lagrangian is real. Their origin is the possible contribution from intermediate on-shell states in the decay process. Since these phases are generated by CP-invariant interactions, they are the same in  $A_f$  and  $\bar{A}_{\bar{f}}$ . Usually the dominant rescattering is due to strong interactions and hence the designation “strong phases” for the phase shifts so induced. Again, only the relative strong phases between different terms in the amplitude are physically meaningful.

The ‘weak’ and ‘strong’ phases discussed here appear in addition to the ‘spurious’ CP-transformation phases of Eq. (B.3). Those spurious phases are due to an arbitrary choice of phase convention, and do

not originate from any dynamics or induce any CP violation. For simplicity, we set them to zero from here on.

It is useful to write each contribution  $a_i$  to  $A_f$  in three parts: its magnitude  $|a_i|$ , its weak phase  $\phi_i$ , and its strong phase  $\delta_i$ . If, for example, there are two such contributions,  $A_f = a_1 + a_2$ , we have

$$\begin{aligned} A_f &= |a_1|e^{i(\delta_1+\phi_1)} + |a_2|e^{i(\delta_2+\phi_2)}, \\ \bar{A}_f &= |a_1|e^{i(\delta_1-\phi_1)} + |a_2|e^{i(\delta_2-\phi_2)}. \end{aligned} \quad (\text{B.22})$$

Similarly, for neutral meson decays, it is useful to write

$$M_{12} = |M_{12}|e^{i\phi_M}, \quad \Gamma_{12} = |\Gamma_{12}|e^{i\phi_\Gamma}. \quad (\text{B.23})$$

Each of the phases appearing in Eqs. (B.22,B.23) is convention dependent, but combinations such as  $\delta_1 - \delta_2$ ,  $\phi_1 - \phi_2$ ,  $\phi_M - \phi_\Gamma$  and  $\phi_M + \phi_1 - \bar{\phi}_1$  (where  $\bar{\phi}_1$  is a weak phase contributing to  $\bar{A}_f$ ) are physical.

In the approximations that only a single weak phase contributes to decay,  $A_f = |a_f|e^{i(\delta_f+\phi_f)}$ , and that  $|\Gamma_{12}/M_{12}| = 0$ , we obtain  $|\lambda_f| = 1$  and the CP asymmetries in decays to a final CP eigenstate  $f$  [Eq. (B.20)] with eigenvalue  $\eta_f = \pm 1$  are given by

$$\mathcal{A}_{fCP}(t) = \mathcal{I}m(\lambda_f) \sin(\Delta mt) \quad \text{with} \quad \mathcal{I}m(\lambda_f) = \eta_f \sin(\phi_M + 2\phi_f). \quad (\text{B.24})$$

Note that the phase so measured is purely a weak phase, and no hadronic parameters are involved in the extraction of its value from  $\mathcal{I}m(\lambda_f)$ .

## C Supersymmetric flavor violation

### C.1 Mass insertions

Supersymmetric models provide, in general, new sources of flavor violation. We here present the formalism of mass insertions. We do that for the charged sleptons, but the formalism is straightforwardly adapted for squarks.

The supersymmetric lepton flavor violation is most commonly analyzed in the basis in which the charged lepton mass matrix and the gaugino vertices are diagonal. In this basis, the slepton masses are not necessarily flavor-diagonal, and have the form

$$\tilde{\ell}_{Mi}^* (M_{\tilde{\ell}}^2)_{ij}^{MN} \tilde{\ell}_{Nj} = (\tilde{\ell}_{Li}^* \tilde{\ell}_{Rk}^*) \begin{pmatrix} M_{Lij}^2 & A_{il}v_d \\ A_{jk}v_d & M_{Rkl}^2 \end{pmatrix} \begin{pmatrix} \tilde{\ell}_{Lj} \\ \tilde{\ell}_{Rl} \end{pmatrix}, \quad (\text{C.1})$$

where  $M, N = L, R$  label chirality, and  $i, j, k, l = 1, 2, 3$  are generational indices.  $M_L^2$  and  $M_R^2$  are the supersymmetry breaking slepton masses-squared. The  $A$  parameters enter in the trilinear scalar couplings  $A_{ij}\phi_d\tilde{\ell}_{Li}\tilde{\ell}_{Rj}^*$ , where  $\phi_d$  is the down-type Higgs boson, and  $v_d = \langle\phi_d\rangle$ . We neglect small flavor-conserving terms involving  $\tan\beta = v_u/v_d$ .

In this basis, charged LFV takes place through one or more slepton mass insertion. Each mass insertion brings with it a factor of

$$\delta_{ij}^{MN} \equiv (M_{\tilde{\ell}}^2)_{ij}^{MN} / \tilde{m}^2, \quad (\text{C.2})$$

where  $\tilde{m}^2$  is the representative slepton mass scale. Physical processes therefore constrain

$$(\delta_{ij}^{MN})_{\text{eff}} \sim \max [\delta_{ij}^{MN}, \delta_{ik}^{MP} \delta_{kj}^{PN}, \dots, (i \leftrightarrow j)]. \quad (\text{C.3})$$

For example,

$$(\delta_{12}^{LR})_{\text{eff}} \sim \max [A_{12}v_d/\tilde{m}^2, M_{L1k}^2 A_{k2}v_d/\tilde{m}^4, A_{1k}v_d M_{Rk2}^2/\tilde{m}^4, \dots, (1 \leftrightarrow 2)]. \quad (\text{C.4})$$

Note that contributions with two or more insertions may be less suppressed than those with only one.

It is useful to express the  $\delta_{ij}^{MN}$  mass insertions in terms of parameters in the mass basis. We can write, for example,

$$\delta_{ij}^{LL} = \frac{1}{\tilde{m}^2} \sum_{\alpha} K_{i\alpha}^L K_{j\alpha}^{L*} \Delta \tilde{m}_{L\alpha}^2. \quad (\text{C.5})$$

Here, we ignore  $L-R$  mixing, so that  $K_{i\alpha}^L$  is the mixing angle in the coupling of a neutralino to  $\ell_{Li} - \tilde{\ell}_{L\alpha}$  (with  $\ell_i = e, \mu, \tau$  denoting charged lepton mass eigenstates and  $\tilde{\ell}_{L\alpha} = \tilde{\ell}_1, \tilde{\ell}_2, \tilde{\ell}_3$  denoting charged slepton mass eigenstates), and  $\Delta \tilde{m}_{L\alpha}^2 = m_{\tilde{\ell}_{L\alpha}}^2 - \tilde{m}^2$ . Using the unitarity of the mixing matrix  $K^L$ , we can write

$$\tilde{m}^2 \delta_{ij}^{LL} = \sum_{\alpha} K_{i\alpha}^L K_{j\alpha}^{L*} (\Delta \tilde{m}_{L\alpha}^2 + \tilde{m}^2) = (M_{\tilde{\ell}}^2)_{ij}^{LL}, \quad (\text{C.6})$$

thus reproducing the definition (C.2).

In many cases, a two generation effective framework is useful. To understand that, consider a case where (no summation over  $i, j, k$ )

$$\begin{aligned} |K_{ik}^L K_{jk}^{L*}| &\ll |K_{ij}^L K_{jj}^{L*}|, \\ |K_{ik}^L K_{jk}^{L*} \Delta m_{\tilde{\ell}_{Lk} \tilde{\ell}_{Li}}^2| &\ll |K_{ij}^L K_{jj}^{L*} \Delta m_{\tilde{\ell}_{Lj} \tilde{\ell}_{Li}}^2|, \end{aligned} \quad (\text{C.7})$$

where  $\Delta m_{\tilde{\ell}_j \tilde{\ell}_i}^2 = m_{\tilde{\ell}_{Lj}}^2 - m_{\tilde{\ell}_{Li}}^2$ . Then, the contribution of the intermediate  $\tilde{\ell}_k$  can be neglected and, furthermore, to a good approximation  $K_{ii}^L K_{ji}^{L*} + K_{ij}^L K_{jj}^{L*} = 0$ . For these cases, we obtain

$$\delta_{ij}^{LL} = \frac{\Delta m_{\tilde{\ell}_{Lj} \tilde{\ell}_{Li}}^2}{\tilde{m}^2} K_{ij}^L K_{jj}^{L*}. \quad (\text{C.8})$$

## C.2 Neutral meson mixing

We consider the squark-gluino box diagram contribution to  $D^0 - \bar{D}^0$  mixing amplitude that is proportional to  $K_{2i}^u K_{1i}^{u*} K_{2j}^u K_{1j}^{u*}$ , where  $K^u$  is the mixing matrix of the gluino couplings to left-handed up quarks and their up squark partners. (In the language of the mass insertion approximation, we calculate here the contribution that is  $\propto [(\delta_{LL}^u)_{12}]^2$ .) We work in the mass basis for both quarks and squarks.

The contribution is given by

$$M_{12}^D = -i \frac{4\pi^2}{27} \alpha_s^2 m_D f_D^2 B_D \eta_{\text{QCD}} \sum_{i,j} (K_{2i}^u K_{1i}^{u*} K_{2j}^u K_{1j}^{u*}) (11 \tilde{I}_{4ij} + 4 \tilde{m}_g^2 I_{4ij}). \quad (\text{C.9})$$

where

$$\begin{aligned} \tilde{I}_{4ij} &\equiv \int \frac{d^4 p}{(2\pi)^4} \frac{p^2}{(p^2 - \tilde{m}_g^2)^2 (p^2 - \tilde{m}_i^2) (p^2 - \tilde{m}_j^2)} \\ &= \frac{i}{(4\pi)^2} \left[ \frac{\tilde{m}_g^2}{(\tilde{m}_i^2 - \tilde{m}_g^2)(\tilde{m}_j^2 - \tilde{m}_g^2)} \right. \\ &\quad \left. + \frac{\tilde{m}_i^4}{(\tilde{m}_i^2 - \tilde{m}_j^2)(\tilde{m}_i^2 - \tilde{m}_g^2)^2} \ln \frac{\tilde{m}_i^2}{\tilde{m}_g^2} + \frac{\tilde{m}_j^4}{(\tilde{m}_j^2 - \tilde{m}_i^2)(\tilde{m}_j^2 - \tilde{m}_g^2)^2} \ln \frac{\tilde{m}_j^2}{\tilde{m}_g^2} \right], \quad (\text{C.10}) \end{aligned}$$

$$I_{4ij} \equiv \int \frac{d^4 p}{(2\pi)^4} \frac{1}{(p^2 - \tilde{m}_g^2)^2 (p^2 - \tilde{m}_i^2) (p^2 - \tilde{m}_j^2)}$$

$$\begin{aligned}
 &= \frac{i}{(4\pi)^2} \left[ \frac{1}{(\tilde{m}_i^2 - \tilde{m}_g^2)(\tilde{m}_j^2 - \tilde{m}_g^2)} \right. \\
 &\quad \left. + \frac{\tilde{m}_i^2}{(\tilde{m}_i^2 - \tilde{m}_j^2)(\tilde{m}_i^2 - \tilde{m}_g^2)^2} \ln \frac{\tilde{m}_i^2}{\tilde{m}_g^2} + \frac{\tilde{m}_j^2}{(\tilde{m}_j^2 - \tilde{m}_i^2)(\tilde{m}_j^2 - \tilde{m}_g^2)^2} \ln \frac{\tilde{m}_j^2}{\tilde{m}_g^2} \right]. \quad (\text{C.11})
 \end{aligned}$$

We now follow the discussion in refs. [21, 80]. To see the consequences of the super-GIM mechanism, let us expand the expression for the box integral around some value  $\tilde{m}_q^2$  for the squark masses-squared:

$$\begin{aligned}
 I_4(\tilde{m}_g^2, \tilde{m}_i^2, \tilde{m}_j^2) &= I_4(\tilde{m}_g^2, \tilde{m}_q^2 + \delta\tilde{m}_i^2, \tilde{m}_q^2 + \delta\tilde{m}_j^2) \\
 &= I_4(\tilde{m}_g^2, \tilde{m}_q^2, \tilde{m}_q^2) + (\delta\tilde{m}_i^2 + \delta\tilde{m}_j^2) I_5(\tilde{m}_g^2, \tilde{m}_q^2, \tilde{m}_q^2, \tilde{m}_q^2) \\
 &\quad + \frac{1}{2} [(\delta\tilde{m}_i^2)^2 + (\delta\tilde{m}_j^2)^2 + 2(\delta\tilde{m}_i^2)(\delta\tilde{m}_j^2)] I_6(\tilde{m}_g^2, \tilde{m}_q^2, \tilde{m}_q^2, \tilde{m}_q^2, \tilde{m}_q^2) + \dots \quad (\text{C.12})
 \end{aligned}$$

where

$$I_n(\tilde{m}_g^2, \tilde{m}_q^2, \dots, \tilde{m}_q^2) \equiv \int \frac{d^4p}{(2\pi)^4} \frac{1}{(p^2 - \tilde{m}_g^2)^2 (p^2 - \tilde{m}_q^2)^{n-2}}, \quad (\text{C.13})$$

and similarly for  $\tilde{I}_{4ij}$ . Note that  $I_n \propto (\tilde{m}_q^2)^{n-2}$  and  $\tilde{I}_n \propto (\tilde{m}_q^2)^{n-3}$ . Thus, using  $x \equiv \tilde{m}_g^2/\tilde{m}_q^2$ , it is customary to define

$$I_n \equiv \frac{i}{(4\pi)^2 (\tilde{m}_q^2)^{n-2}} f_n(x), \quad \tilde{I}_n \equiv \frac{i}{(4\pi)^2 (\tilde{m}_q^2)^{n-3}} \tilde{f}_n(x). \quad (\text{C.14})$$

The unitarity of the mixing matrix implies that

$$\sum_i (K_{2i}^u K_{1i}^{u*} K_{2j}^u K_{1j}^{u*}) = \sum_j (K_{2i}^u K_{1i}^{u*} K_{2j}^u K_{1j}^{u*}) = 0. \quad (\text{C.15})$$

Consequently, the terms that are proportional  $f_4, \tilde{f}_4, f_5$  and  $\tilde{f}_5$  vanish in their contribution to  $M_{12}$ . When  $\delta\tilde{m}_i^2 \ll \tilde{m}_q^2$  for all  $i$ , the leading contributions to  $M_{12}$  come from  $f_6$  and  $\tilde{f}_6$ . We learn that for quasi-degenerate squarks, the leading contribution is quadratic in the small mass-squared difference. The functions  $f_6(x)$  and  $\tilde{f}_6(x)$  are given by

$$\begin{aligned}
 f_6(x) &= \frac{6(1+3x)\ln x + x^3 - 9x^2 - 9x + 17}{6(1-x)^5}, \\
 \tilde{f}_6(x) &= \frac{6x(1+x)\ln x - x^3 - 9x^2 + 9x + 1}{3(1-x)^5}. \quad (\text{C.16})
 \end{aligned}$$

For example, with  $x = 1$ ,  $f_6(1) = -1/20$  and  $\tilde{f}_6 = +1/30$ ; with  $x = 2.33$ ,  $f_6(2.33) = -0.015$  and  $\tilde{f}_6 = +0.013$ .

To further simplify things, let us consider a two generation case. Then

$$\begin{aligned}
 M_{12}^D &\propto 2(K_{21}^u K_{11}^{u*})^2 (\delta\tilde{m}_1^2)^2 + 2(K_{22}^u K_{12}^{u*})^2 (\delta\tilde{m}_2^2)^2 + (K_{21}^u K_{11}^{u*} K_{22}^u K_{12}^{u*}) (\delta\tilde{m}_1^2 + \delta\tilde{m}_2^2)^2 \\
 &= (K_{21}^u K_{11}^{u*})^2 (\tilde{m}_2^2 - \tilde{m}_1^2)^2. \quad (\text{C.17})
 \end{aligned}$$

We thus rewrite Eq. (C.9) for the case of quasi-degenerate squarks:

$$M_{12}^D = \frac{\alpha_s^2 m_D f_D^2 B_D \eta_{\text{QCD}}}{108 \tilde{m}_q^2} [11\tilde{f}_6(x) + 4x f_6(x)] \frac{(\Delta\tilde{m}_{21}^2)^2}{\tilde{m}_q^4} (K_{21}^u K_{11}^{u*})^2. \quad (\text{C.18})$$

For example, for  $x = 1$ ,  $11\tilde{f}_6(x) + 4x f_6(x) = +0.17$ . For  $x = 2.33$ ,  $11\tilde{f}_6(x) + 4x f_6(x) = +0.003$ .

## Acknowledgements

I thank my students – Yonit Hochberg, Daniel Grossman, Aielet Efrati and Avital Dery – for many useful discussions. The research of Y.N. is supported by the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation (grant No 1937/12), by the Israel Science Foundation (grant No 579/11), and by the German–Israeli Foundation (GIF) (Grant No G-1047-92.7/2009).

## References

- [1] M. Kobayashi and T. Maskawa, *Prog. Theor. Phys.* **49** (1973) 652.
- [2] N. Cabibbo, *Phys. Rev. Lett.* **10** (1963) 531.
- [3] A. B. Carter and A. I. Sanda, *Phys. Rev. Lett.* **45** (1980) 952; *Phys. Rev. D* **23** (1981) 1567.
- [4] I. I. Y. Bigi and A. I. Sanda, *Nucl. Phys. B* **193** (1981) 85.
- [5] G. Buchalla, A. J. Buras, and M. E. Lautenbacher, *Rev. Mod. Phys.* **68** (1996) 1125 [arXiv:hep-ph/9512380].
- [6] Y. Grossman, A. L. Kagan and Z. Ligeti, *Phys. Lett. B* **538** (2002) 327 [arXiv:hep-ph/0204212].
- [7] H. Boos, T. Mannel and J. Reuter, *Phys. Rev. D* **70** (2004) 036006 [arXiv:hep-ph/0403085].
- [8] H. n. Li and S. Mishima, *JHEP* **0703** (2007) 009 [arXiv:hep-ph/0610120].
- [9] M. Gronau and J. L. Rosner, *Phys. Lett. B* **672** (2009) 349 [arXiv:0812.4796 [hep-ph]].
- [10] Y. Amhis *et al.* [Heavy Flavor Averaging Group Collaboration], arXiv:1207.1158 [hep-ex] and online update at <http://www.slac.stanford.edu/xorg/hfag>.
- [11] J. Beringer *et al.* [Particle Data Group Collaboration], *Phys. Rev. D* **86** (2012) 010001.
- [12] CKMfitter Group (J. Charles *et al.*), *Eur. Phys. J.* **C41** (2005) 1 [hep-ph/0406184], updated results and plots available at: <http://ckmfitter.in2p3.fr>
- [13] Y. Nir, *Nucl. Phys. Proc. Suppl.* **117** (2003) 111 [arXiv:hep-ph/0208080].
- [14] Y. Grossman, Y. Nir and M. P. Worah, *Phys. Lett. B* **407** (1997) 307 (1997).
- [15] Y. Grossman, Y. Nir and G. Raz, *Phys. Rev. Lett.* **97** (2006) 151801 [arXiv:hep-ph/0605028].
- [16] M. Bona *et al.* [UTfit Collaboration], *JHEP* **0803** (2008) 049 [arXiv:0707.0636 [hep-ph]].
- [17] G. C. Branco, L. Lavoura and J. P. Silva, *CP violation*, Clarendon Press, Oxford (1999).
- [18] I. I. Y. Bigi and N. G. Uraltsev, *Nucl. Phys. B* **592** (2001) 92 [arXiv:hep-ph/0005089].
- [19] A. F. Falk, Y. Grossman, Z. Ligeti and A. A. Petrov, *Phys. Rev. D* **65** (2002) 054034 [arXiv:hep-ph/0110317].
- [20] A. F. Falk, Y. Grossman, Z. Ligeti, Y. Nir and A. A. Petrov, *Phys. Rev. D* **69** (2004) 114021 [arXiv:hep-ph/0402204].
- [21] G. Raz, *Phys. Rev. D* **66** (2002) 037701 [arXiv:hep-ph/0205310].
- [22] G. Isidori, Y. Nir and G. Perez, *Ann. Rev. Nucl. Part. Sci.* **60** (2010) 355 [arXiv:1002.0900 [hep-ph]].
- [23] N. Arkani-Hamed and S. Dimopoulos, *JHEP* **0506** (2005) 073 [arXiv:hep-th/0405159].
- [24] A. G. Cohen, D. B. Kaplan and A. E. Nelson, *Phys. Lett. B* **388** (1996) 588 [arXiv:hep-ph/9607394].
- [25] Y. Nir and N. Seiberg, *Phys. Lett. B* **309** (1993) 337 [arXiv:hep-ph/9304307].
- [26] M. Leurer, Y. Nir and N. Seiberg, *Nucl. Phys. B* **420** (1994) 468 [arXiv:hep-ph/9310320].
- [27] M. Ciuchini, E. Franco, D. Guadagnoli, V. Lubicz, M. Pierini, V. Porretti and L. Silvestrini, *Phys. Lett. B* **655** (2007) 162 [arXiv:hep-ph/0703204].
- [28] Y. Nir, *JHEP* **0705** (2007) 102 [arXiv:hep-ph/0703235].
- [29] O. Gedalia, J. F. Kamenik, Z. Ligeti and G. Perez, *Phys. Lett. B* **714** (2012) 55 [arXiv:1202.5038 [hep-ph]].

- [30] K. Blum, Y. Grossman, Y. Nir and G. Perez, *Phys. Rev. Lett.* **102** (2009) 211802 [arXiv:0903.2118 [hep-ph]].
- [31] G. D'Ambrosio, G. F. Giudice, G. Isidori and A. Strumia, *Nucl. Phys. B* **645** (2002) 155 [arXiv:hep-ph/0207036].
- [32] Y. Grossman, Y. Nir, J. Thaler, T. Volansky and J. Zupan, *Phys. Rev. D* **76** (2007) 096006 [arXiv:0706.1845 [hep-ph]].
- [33] V. Cirigliano, B. Grinstein, G. Isidori and M. B. Wise, *Nucl. Phys. B* **728** (2005) 121 [arXiv:hep-ph/0507001].
- [34] V. Cirigliano and B. Grinstein, *Nucl. Phys. B* **752** (2006) 18 [arXiv:hep-ph/0601111].
- [35] V. Cirigliano, G. Isidori and V. Porretti, *Nucl. Phys. B* **763** (2007) 228 [arXiv:hep-ph/0607068].
- [36] G. C. Branco, A. J. Buras, S. Jager, S. Uhlig and A. Weiler, *JHEP* **0709** (2007) 004 [arXiv:hep-ph/0609067].
- [37] M. C. Chen and H. B. Yu, *Phys. Lett. B* **672** (2009) 253 [arXiv:0804.2503 [hep-ph]].
- [38] E. Gross, D. Grossman, Y. Nir and O. Vitells, *Phys. Rev. D* **81** (2010) 055013 [arXiv:1001.2883 [hep-ph]].
- [39] J. L. Feng, C. G. Lester, Y. Nir and Y. Shadmi, *Phys. Rev. D* **77** (2008) 076002 [arXiv:0712.0674 [hep-ph]].
- [40] J. L. Feng, I. Galon, D. Sanford, Y. Shadmi and F. Yu, *Phys. Rev. D* **79** (2009) 116009 [arXiv:0904.1416 [hep-ph]].
- [41] J. L. Feng, S. T. French, C. G. Lester, Y. Nir and Y. Shadmi, *Phys. Rev. D* **80** (2009) 114004 [arXiv:0906.4215 [hep-ph]].
- [42] J. L. Feng *et al.*, *JHEP* **1001** (2010) 047 [arXiv:0910.1618 [hep-ph]].
- [43] G. Hiller and Y. Nir, *JHEP* **0803** (2008) 046 [arXiv:0802.0916 [hep-ph]].
- [44] G. Hiller, Y. Hochberg and Y. Nir, *JHEP* **0903** (2009) 115 [arXiv:0812.0511 [hep-ph]].
- [45] G. Hiller, Y. Hochberg and Y. Nir, *JHEP* **1003** (2010) 079 [arXiv:1001.1513 [hep-ph]].
- [46] C. D. Froggatt and H. B. Nielsen, *Nucl. Phys. B* **147** (1979) 277.
- [47] M. Dine, R. G. Leigh and A. Kagan, *Phys. Rev. D* **48** (1993) 4269 [hep-ph/9304299].
- [48] A. E. Nelson and M. J. Strassler, *JHEP* **0009** (2000) 030 [arXiv:hep-ph/0006251];
- [49] N. Arkani-Hamed and M. Schmaltz, *Phys. Rev. D* **61** (2000) 033005 [hep-ph/9903417].
- [50] M. C. Gonzalez-Garcia, M. Maltoni, J. Salvado and T. Schwetz, *JHEP* **1212** (2012) 123 [arXiv:1209.3023 [hep-ph]].
- [51] L. J. Hall, H. Murayama and N. Weiner, *Phys. Rev. Lett.* **84** (2000) 2572 [hep-ph/9911341].
- [52] N. Haba and H. Murayama, *Phys. Rev. D* **63** (2001) 053010 [hep-ph/0009174].
- [53] A. de Gouvea and H. Murayama, *Phys. Lett. B* **573** (2003) 94 [hep-ph/0301050]; arXiv:1204.1249 [hep-ph].
- [54] S. Amitai, arXiv:1211.6252 [hep-ph].
- [55] P. F. Harrison, D. H. Perkins and W. G. Scott, *Phys. Lett. B* **530** (2002) 167 [hep-ph/0202074].
- [56] E. Ma and G. Rajasekaran, *Phys. Rev. D* **64** (2001) 113012 [hep-ph/0106291].
- [57] G. Altarelli and F. Feruglio, *Rev. Mod. Phys.* **82** (2010) 2701 [arXiv:1002.0211 [hep-ph]]. *Nucl. Phys. B* **741** (2006) 215 [hep-ph/0512103].
- [58] F. P. An *et al.* [DAYA-BAY Collaboration], *Phys. Rev. Lett.* **108** (2012) 171803 [arXiv:1203.1669 [hep-ex]].
- [59] S. Amitai, arXiv:1212.5165 [hep-ph].
- [60] G. Aad *et al.* [ATLAS Collaboration], *Phys. Lett. B* **716** (2012) 1 [arXiv:1207.7214 [hep-ex]].
- [61] S. Chatrchyan *et al.* [CMS Collaboration], *Phys. Lett. B* **716** (2012) 30 [arXiv:1207.7235 [hep-ex]].

- [62] D. Carmi, A. Falkowski, E. Kuflik and T. Volansky, arXiv:1206.4201 [hep-ph].
- [63] A. Dery, A. Efrati, Y. Hochberg and Y. Nir, *JHEP* **1305** (2013) 039 [arXiv:1302.3229 [hep-ph]].
- [64] A. Dery, A. Efrati, G. Hiller, Y. Hochberg and Y. Nir, arXiv:1304.6727 [hep-ph].
- [65] G. Blankenburg, J. Ellis and G. Isidori, *Phys. Lett. B* **712** (2012) 386 [arXiv:1202.5704 [hep-ph]].
- [66] R. Harnik, J. Kopp and J. Zupan, *JHEP* **1303** (2013) 026 [arXiv:1209.1397 [hep-ph]].
- [67] S. Davidson and P. Verdier, *Phys. Rev. D* **86** (2012) 111701 [arXiv:1211.1248 [hep-ph]].
- [68] A. Arhrib, Y. Cheng and O. C. W. Kong, *Phys. Rev. D* **87** (2013) 015025 [arXiv:1210.8241 [hep-ph]].
- [69] Y. Grossman and Y. Nir, *Nucl. Phys. B* **448** (1995) 30 [hep-ph/9502418].
- [70] L. Chau and W. Keung, *Phys. Rev. Lett.* **53** (1984) 1802.
- [71] L. Wolfenstein, *Phys. Rev. Lett.* **51** (1983) 1945.
- [72] A. J. Buras, M. E. Lautenbacher, and G. Ostermaier, *Phys. Rev. D* **50** (1994) 3433 [arXiv:hep-ph/9403384].
- [73] C. Dib, I. Dunietz, F. J. Gilman and Y. Nir, *Phys. Rev. D* **41** (1990) 1522.
- [74] J. L. Rosner, A. I. Sanda and M. P. Schmidt, EFI-88-12-CHICAGO [Presented at Workshop on High Sensitivity Beauty Physics, Batavia, IL, Nov 11-14, 1987].
- [75] V. Weisskopf and E. P. Wigner, *Z. Phys.* **63** (1930) 54; *Z. Phys.* **65** (1930) 18. [See Appendix A of P. K. Kabir, “The CP Puzzle: Strange Decays of the Neutral Kaon”, Academic Press (1968).]
- [76] Y. Nir, SLAC-PUB-5874 [Lectures given at 20th Annual SLAC Summer Institute on Particle Physics (Stanford, CA, 1992)].
- [77] I. Dunietz and J. L. Rosner, *Phys. Rev. D* **34** (1986) 1404.
- [78] Ya. I. Azimov, N. G. Uraltsev, and V. A. Khoze, *Sov. J. Nucl. Phys.* **45** (1987) 878 [*Yad. Fiz.* **45** (1987) 1412].
- [79] I. I. Bigi and A. I. Sanda, *Nucl. Phys. B* **281** (1987) 41.
- [80] Y. Nir and G. Raz, *Phys. Rev. D* **66** (2002) 035007 [arXiv:hep-ph/0206064].

## Bibliography

G.C. Branco, L. Lavoura and J.P. Silva, *CP Violation* (Oxford University Press, Oxford, 1999).  
 H.R. Quinn and Y. Nir, *The Mystery of the Missing Antimatter* (Princeton University Press, Princeton, 2007).



## QCD under extreme conditions: an informal discussion

*E.S. Fraga\**

J. W. Goethe-University, Frankfurt am Main, Germany

### Abstract

We present an informal discussion of some aspects of strong interactions under extreme conditions of temperature and density at an elementary level. This summarizes lectures delivered at the 2013 CERN – Latin-American School of High-Energy Physics and is aimed at students working in experimental high-energy physics.

### 1 Introduction and motivation: why, where and how

Quantum Chromodynamics (QCD) is an extremely successful theory of strong interactions that has passed numerous tests in particle accelerators over more than 40 years [1]. This corresponds to the behavior of hadrons in the vacuum, including not only the spectrum but also all sorts of dynamical processes. More recently strong interactions, and therefore QCD, has also started being probed in a medium, under conditions that become more and more extreme [2]. Although quite involved theoretically, this is not just an academic problem. In order to make it clear, one should consider three very basic questions, that should always be asked in the beginning: why? where? how?

#### 1.1 Why?

It was realized since the very beginning that strong interactions exhibit two remarkable features that are related but represent properties of complementary sectors of the energy scale. The first one is asymptotic freedom [3], which can be perturbatively demonstrated by an explicit computation of the beta function to a give loop order in QCD [4]. The second, which is consistent with the first but should be seen as totally independent, since it is a property of the nonperturbative vacuum of strong interactions, is color confinement [5]. Even though reality constantly shows that confinement is a property of strong interactions, and therefore should somehow be built in QCD, this proof remains a theoretical open problem so far. Even for the pure Yang-Mills theory, where the bound states correspond to glueballs, the existence of a mass gap is still to be shown after more than half a century of the original paper on nonabelian gauge theories [6]. For this reason, confinement is ranked in the Clay Mathematics Institute list of unsolved Millennium problems [7].

Much more than a cute (and very tough) mathematical problem, this is certainly among the most important theoretical and phenomenological problems in particle physics, since hidden there is the real origin of mass, as we feel in our everyday lives and experience with ordinary (and not so ordinary) matter. Although the Higgs mechanism provides a way to give mass to elementary particles in the Standard Model [8], most of what constitutes the masses of hadrons come from interactions. For instance, more than 90% of the proton mass originates in quark and gluon condensates [9]. So, in spite of the fantastic success of the Standard Model [8], we do not understand a few essential mechanisms.

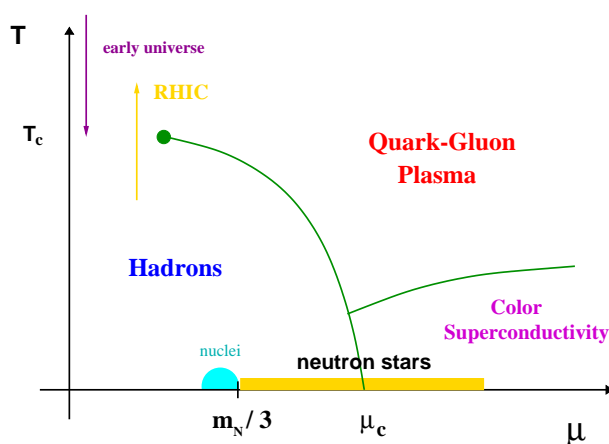
Extremely high temperatures and densities bring us to an energy scale that facilitates deconfinement, and matter under such extreme conditions can behave in unexpected ways due to collective effects. This is, of course, a way to study the mechanism of confinement (by perturbing or modifying this state of matter). This leads us also to a deeper yet childish motivation, that of understanding what happens if we keep making things hotter and hotter, or keep squeezing things harder and harder [10]. These questions can be reformulated in a more technical fashion as 'what is the inner structure of matter and the nature

---

\*On leave from Instituto de Física, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil.

of strong interactions under extreme conditions of temperature and density?'. In experiments, one needs to “squeeze”, “heat” and “break”. From the theoretical point of view, one needs a good formulation of in-medium quantum field theory, using QCD or effective theories.

It is clear that the challenge is enormous. Although confinement seems to be a key feature of hadrons, and manifests also in relevant scales such as  $f_\pi$  or  $\Lambda_{QCD}$ , it only *seems* to be present in QCD. So far, controlled lattice simulations show strong evidence of confinement in the pure gauge theory [11]. As hinted previously, however, the theory is nonperturbative at the relevant scales, so that analytic methods are very constrained. And, although lattice simulations have developed to provide solid results in several scenarios, they are not perfect. And, more important, they are not Nature. To make progress in understanding, or at least collecting important facts, one needs it all: experiments and observations, lattice simulations, the full theory in specific (solvable to some extent) limits and effective models. And also combinations, whenever possible, to diminish the drawbacks of each approach.



**Fig. 1:** Cartoon of a phase diagram for strong interactions. Extracted from Ref. [12]

Whichever the framework chosen, collective phenomena will play a major role. Although somewhat put aside in the so-called microscopic “fundamental” particle physics, collective effects can affect dramatically the behavior of elementary particles in a medium under certain conditions. Besides the well-known examples of BCS and BEC phases in condensed matter systems [13], and also in dense quark matter [14], it was recently found that photons can form a Bose-Einstein condensate [15]. In fact, the textbook case of water and its different phases is quite illustrative of the richness that comes from collective phenomena that would hardly be guessed from the case of very few or non-interacting elementary particles.

In terms of the thermodynamics, or many-body problem, the basic idea is to perturb the (confined) vacuum to study confinement by heating (temperature), squeezing or unbalancing species (chemical potentials for baryon number, isospin, strangeness, etc) and using classical external fields (magnetic, electric, etc), so that the system is taken away from the confined phase and back. One can also relate (or not) confinement to other key properties of strong interactions, such as chiral symmetry. And, from the theorist standpoint, draw all possible phase diagrams of QCD and its “cousin theories” (realizations of QCD with parameters, such as the number of colors or flavors, or the values of masses, that are not realized in Nature) to learn basic facts. There are several examples, one well-known being the ‘Columbia plot’, where one studies the nature of the phase transitions and critical lines on the  $(m_u = m_d, m_s)$  plane. Nevertheless, if one draws a cartoon of the phase diagram in the temperature vs. quark chemical potential, for instance Fig. 1, and compares it to computations from effective models, lattice simulations and freeze-out points extracted from high-energy heavy ion collision data, one sees that the points still scatter in a large area [16]. So, there is still a long way ahead.

## 1.2 Where?

According to the Big Bang picture and the current description of the evolution of the early universe [17], we expect that at about  $10^{-5}s$  after the Big Bang a soup of quark-gluon plasma (in the presence of electrons, photons, etc) has undergone a phase transition to confined hadrons. This was, of course, the first realization of a QCD transition. This process was thermally driven and happened at very low baryon chemical potential.

It is quite remarkable that the scales of strong interactions allow for the experimental reproduction of analogous conditions in high-energy ultra-relativistic heavy ion collisions in the laboratory [18]. In a picture by T. D. Lee, these collisions are seen as heavy bulls that collide and generate new states of matter [19]. Such experiments are under way at BNL-RHIC [20] and CERN-LHC [21], and will be part of the future heavy ion programs at FAIR-GSI [22] and NICA [23].

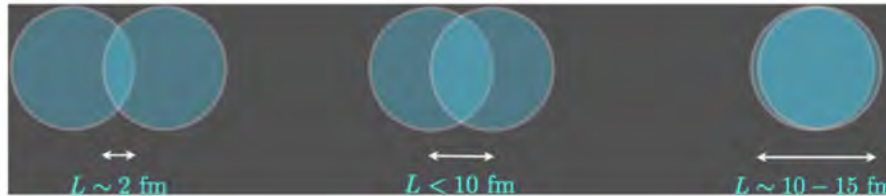
For obvious reasons, it is common to refer to such experiments as ‘‘Little Bangs’’. However, one should be cautious with this point. In spite of the fact that the typical energy scales involved need to be the same, as well as the state of matter created, the so-called quark-gluon plasma [24], the relevant space-time scales differ by several orders of magnitude. Using a simple approximation for the equation of state,

$$3p \approx \epsilon \approx \frac{\pi^2}{30} N(T) T^4, \quad (1)$$

where  $p$  is the pressure,  $\epsilon$  the energy density and  $N(T)$  the number of relevant degrees of freedom, we can easily estimate the typical sizes involved. The radius of the universe at the QCD phase transition epoch, as given by the particle horizon in a Robertson-Walker space-time [25], where the scale factor grows as  $a(t) \sim t^n$ , is given by ( $n = 1/2$  and  $N(T) \sim 50$  at this time for QCD)

$$L_{\text{univ}}(T) \approx \frac{1}{4\pi} \left( \frac{1}{1-n} \right) \left( \frac{45}{\pi N(T)} \right)^{1/2} \frac{M_{\text{Pl}}}{T^2} = \frac{1.45 \times 10^{18}}{(T/\text{GeV})^2 \sqrt{N(T)}} \text{fm}. \quad (2)$$

Here  $M_{\text{Pl}}$  is the Planck mass, and it is clear that the system is essentially in the thermodynamic limit.



**Fig. 2:** Cartoon representing non-central heavy ion collisions and how they affect the size of the system.

On the other hand, in heavy ion collisions the typical length scale of the system is  $L_{\text{QGP}} \lesssim 10 - 15 \text{ fm}$ , so that the system can be very small, especially if one considers non-central collisions [26] (see Fig. 2). One can develop analogous arguments for the time scales given by the expansion rates, finding that the whole process in the early universe happens adiabatically, whereas in heavy ions it is not even clear whether the system can achieve thermal equilibrium, given the explosive nature of the evolution in this case. So, there are certainly large differences (in time and length scales) between Big and Little Bangs...

Keeping this caveat in mind, heavy ion experiments have been investigating new phases of matter at very high energies for more than a decade, producing an awesome amount of interesting data and a richer picture of strong interactions (see Ref. [27] for a review).

In the realization of the Big and Little Bangs one is always in the high temperature and low density (small baryon chemical potential) sector of the phase diagram of strong interactions. However,

high densities (at very low temperatures) can also probe new states of hadronic matter, and that is what is expected to be found in the core of compact stars [28]. There, new phases, condensates and even color superconductivity may be present. In particular, the deconfinement and chiral transitions might affect significantly the explosion mechanism in supernovae [28] via modifications in the equation of state.

After a neutron (or hybrid) star is formed, densities in its core can in principle reach several times the nuclear saturation density  $n_0 = 0.16 \text{ fm}^{-3} = 3 \times 10^{14} \text{ g/cm}^3$ , which corresponds to squeezing  $\sim 2$  solar masses into a sphere of  $\sim 10 \text{ km}$  of radius. To describe these objects, one needs General Relativity besides in-medium quantum field theory.

### 1.3 How?

The reader is hopefully already convinced that, in order to describe the phenomenology of the phase structure and dynamics of strong interactions under extreme conditions, one needs all possibilities at disposal: theory, effective modeling, etc. We do not have one problem ahead, but a myriad of different problems. So, one has to make a choice. Our focus here will be the equation of state, of which we will discuss a few aspects.

At this point, we are lead again to the “why” question. And the answer is because, besides carrying all the thermodynamic equilibrium information we may be interested in, it is also the basic crucial ingredient for dynamics, structure, etc. In fact, the phase diagram topology is determined in every detail by the full knowledge of the pressure  $p(T, \mu, B, \dots)$ . This will determine all phases present as we dial different knobs, or control parameters, such as temperature or chemical potentials.

The structure of a compact star, for instance, is given by the solution of the Tolman-Oppenheimer-Volkov (TOV) equations [28], which encode Einstein’s General Relativity field equations in hydrostatic equilibrium for a spherical geometry:

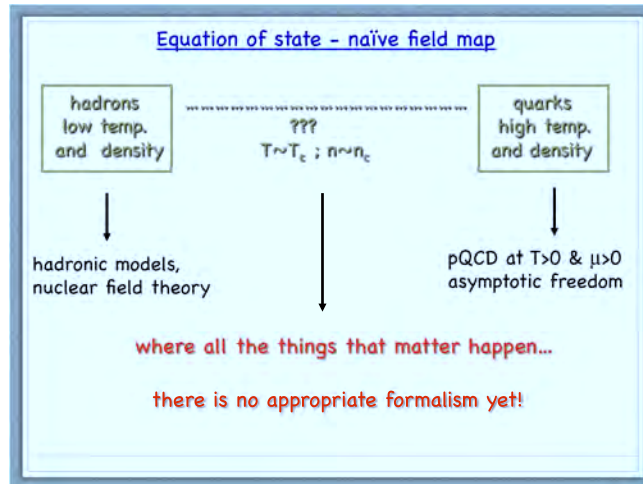
$$\frac{dp}{dr} = -\frac{GM(r)\epsilon(r)}{r^2 \left[1 - \frac{2GM(r)}{r}\right]} \left[1 + \frac{p(r)}{\epsilon(r)}\right] \left[1 + \frac{4\pi r^3 p(r)}{\mathcal{M}(r)}\right], \quad (3)$$

$$\frac{d\mathcal{M}}{dr} = 4\pi r^2 \epsilon(r) ; \quad \mathcal{M}(R) = M. \quad (4)$$

Given the equation of state  $p = p(\epsilon)$ , one can integrate the TOV equations from the origin until the pressure vanishes,  $p(R) = 0$ . Different equations of state define different types of stars (white dwarfs, neutron stars, strange stars, quark stars, etc) and curves on the mass-radius diagram for the families of stars.

Furthermore, to describe the evolution of the hot plasma created in high-energy heavy ion collisions, one need to make use of hydrodynamics, whose fundamental equations encode the conservation of energy-momentum ( $\partial_\mu T^{\mu\nu} = 0$ ) and of baryon number (or different charges) ( $\partial_\mu n_B v^\mu = 0$ , with  $v^\mu v_\mu = 1$ ). These represent only five equations for six unknown functions, the additional constraint provided by the equation of state. Hence, it is clear that we really need the equation of state to make any progress.

In principle, we have all the building blocks to compute the equation of state. The Lagrangian of QCD is given, so one would have “simply” to compute the thermodynamic potential, from which one can extract all relevant thermodynamic functions. The fact that the vacuum of QCD is highly nonperturbative, as discussed previously, makes it way more complicated from the outset. As we know, QCD matter becomes simpler at very high temperatures and densities,  $T$  and  $\mu$  playing the role of the momentum scale in a plasma, but very complicated in the opposite limit. On top of that,  $T$  and  $\mu$  are, unfortunately, not high enough in the interesting cases, so that the physically relevant region is way before asymptotic freedom really kicks in. Perturbative calculations are still an option, but then one has to recall that finite-temperature perturbative QCD is very sick in the infrared, and its naïve formulation breaks down at a



**Fig. 3:** Cartoon of the naïve field map for the equation of state for strong interactions.

scale given by  $g^2T$  [29]. This is known as Linde’s problem: at this scale, for a  $(\ell + 1)$ -loop diagram for the pressure, for  $\ell > 3$  all loops contribute to the term of order  $g^6$  even for weak coupling [29].

The situation does not look very promising, as illustrated by the cartoon of Fig. 3 which shows that there is no appropriate formalism to tackle with the problem in the physically relevant region for the phase structure, namely the critical regions. However, there are several ways out. Some popular examples being: very intelligent and sophisticated “brute force” (lattice QCD), intensive use of symmetries (effective field theory models), redefining degrees of freedom (quasiparticle models), “moving down” from very high-energy perturbative QCD, “moving up” from hadronic low-energy (nuclear) models. And we can and should also combine these possibilities, as discussed previously.

## 2 Symmetries of QCD and effective model building

### 2.1 The simplest approach: the bag model

Before discussing the building of effective models based on the symmetries, or rather approximate symmetries, of QCD, let us consider a very simple description: the MIT bag model [29] applied to describe the thermodynamics of strong interactions.

The model incorporates two basic ingredients, asymptotic freedom and confinement, in the simplest and crudest fashion: bubbles (bags) of perturbative vacuum in a confining medium, including eventual  $O(\alpha_s)$  corrections. Asymptotic freedom is implemented by considering free quarks and gluons inside color singlet bags, whereas confinement is realized by imposing that the vector current vanishes on the boundary.

Then, confinement is achieved by assuming a constant energy density for the vacuum (negative pressure), encoded in the so-called bag constant  $B$ , a phenomenological parameter extracted from fits to hadron masses.  $B$  can also be viewed as the difference in energy density between the QCD and the perturbative vacua. A hadron energy (for a spherical bag) receives contributions from the vacuum and the kinetic energy, so that its minimum yields

$$E_h^{\min} = \frac{16}{3}\pi R_h^3 B, \tag{5}$$

and the hadron pressure (at equilibrium)

$$p_h = \frac{\partial E_h}{\partial V} = -B + \frac{\text{const}}{4\pi R^4} = 0. \quad (6)$$

Assuming the existence of a deconfining transition, the pressure in the quark-gluon plasma phase within this model is given by

$$p_{\text{QGP}} = \left( \nu_b + \frac{7}{4}\nu_f \right) \frac{\pi^2 T^4}{90} - B, \quad (7)$$

whereas the pressure in the hadronic phase (taking, for simplicity, a pion gas) is given by

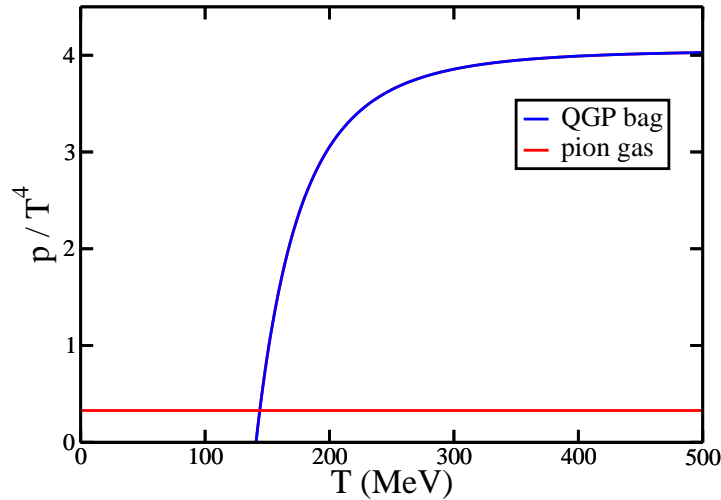
$$p_\pi = \nu_\pi \frac{\pi^2 T^4}{90}, \quad (8)$$

neglecting masses. Here, we have the following numbers of degrees of freedom:  $\nu_\pi = 3$ ,  $\nu_b = 2(N_c^2 - 1)$  and  $\nu_f = 2N_c N_f$  for pions, gluons and quarks, respectively.

For instance, for  $N_c = 3$ ,  $N_f = 2$  and  $B^{1/4} = 200$  MeV, we obtain the following critical temperature:

$$T_c = \left( \frac{45B}{17\pi^2} \right)^{1/4} \approx 144 \text{ MeV} \quad (9)$$

and a first-order phase transition as is clear from Fig. 4. The value of the critical temperature is actually very good as compared to recent lattice simulations [30], considering that this is a very crude model. On the other hand the nature of the transition, a crossover, is almost by construction missed in this approach.



**Fig. 4:** Pressures in the bag model description.

## 2.2 Basics of effective model building in QCD

To go beyond in the study of the phases of QCD, one needs to know its symmetries, and how they are broken spontaneously or explicitly. But QCD is very involved. First, it is a non-abelian  $SU(N_c)$  gauge theory, with gluons living in the adjoint representation. Then, there are  $N_f$  dynamical quarks who live in the fundamental representation. On top of that, these quarks have masses which are all different, which is very annoying from the point of view of symmetries. So, in studying the phases of QCD, we should

do it by parts, and consider many “cousin theories” which are very similar to QCD but simpler (more symmetric). In so doing, we can also study the dependence of physics on parameters which are fixed in Nature.

Fig. 5 illustrates the step-by-step process one can follow in assembling the symmetry features present in QCD and learning from simpler theories, as well as cousin theories. Notice that the full theory, whose parameters are given by comparison to the experimental measurements, has essentially no symmetry left. Yet, some symmetries are mildly broken so that a “memory” of them remains. This fact allows us to use “approximate order parameters”, for instance, a concept that is very useful in practice to characterize the chiral and deconfinement transitions.

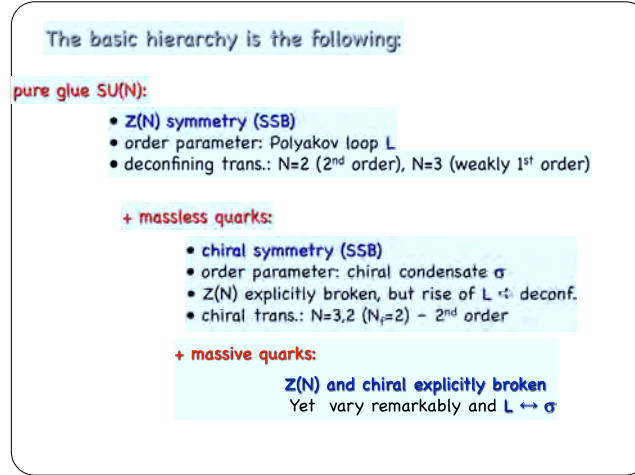


Fig. 5: Basic hierarchy in the step-by-step approach to QCD.

### 2.3 $SU(N_c)$ , $Z(N_c)$ and the Polyakov loop

In the QCD Lagrangian with massless quarks,

$$\mathcal{L} = \frac{1}{2} \text{Tr} F_{\mu\nu} F^{\mu\nu} + \bar{q} i \gamma^\mu D_\mu q, \quad (10)$$

$$D_\mu \equiv (\partial_\mu - ig A_\mu), \quad (11)$$

$$F_{\mu\nu} = \frac{i}{g} [D_\mu(A), D_\nu(A)], \quad (12)$$

we have invariance under local  $SU(N_c)$ . In particular, we have invariance under elements of the center group  $Z(N_c)$  (for a review, see Ref. [31])

$$\Omega_c = e^{i \frac{2n\pi}{N_c} \mathbf{1}}. \quad (13)$$

At finite temperature, one has also to impose the following boundary conditions:

$$A_\mu(\vec{x}, \beta) = +A_\mu(\vec{x}, 0), \quad (14)$$

$$q(\vec{x}, \beta) = -q(\vec{x}, 0). \quad (15)$$

Any gauge transformation that is periodic in  $\tau$  will do it. However, 't Hooft noticed that the class of possible transformations is more general. They are such that

$$\Omega(\vec{x}, \beta) = \Omega_c, \quad \Omega(\vec{x}, 0) = \mathbf{1}, \quad (16)$$

keeping the gauge fields invariant but not the quarks.

For pure glue this  $Z(N_c)$  symmetry is exact and we can define an order parameter - the Polyakov loop:

$$L(\vec{x}) = \frac{1}{N_c} \text{Tr} \mathcal{P} \exp \left[ ig \int_0^\beta d\tau \tau^a A_0^a(\vec{x}, \tau) \right], \quad (17)$$

with  $L$  transforming as

$$L(\vec{x}) \mapsto \Omega_c L(\vec{x}) \mathbf{1} = e^{i\frac{2n\pi}{N_c}} L(\vec{x}). \quad (18)$$

At very high temperatures,  $g \sim 0$ , and  $\beta \mapsto 0$ , so that

$$\langle \ell \rangle = e^{i\frac{2n\pi}{N_c}} \ell_0, \quad \ell_0 \sim 1, \quad (19)$$

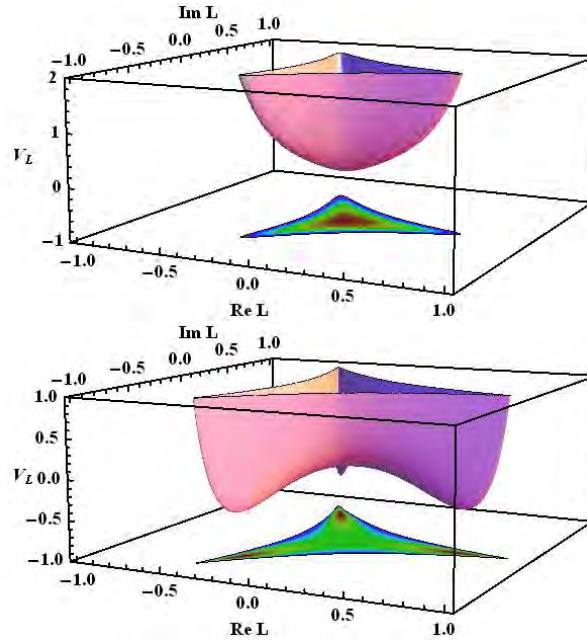
and we have a  $N$ -fold degenerate vacuum, signaling spontaneous symmetry breaking of global  $Z(N_c)$ . At  $T = 0$ , confinement implies that  $\ell_0 = 0$ . Then,  $\ell_0 = 0$  can be used as an order parameter for the deconfining transition:

$$\ell_0 = 0, \quad T < T_c \quad ; \quad \ell_0 > 0, \quad T > T_c. \quad (20)$$

Usually the Polyakov loop is related to the free energy of an infinitely heavy test quark via (confinement: no free quark)

$$\langle \ell \rangle = e^{-F_{test}/T}. \quad (21)$$

See, however, the critical discussion in Ref. [31].



**Fig. 6:** Effective potential for the Polyakov loop for  $T < T_c$  (upper) and  $T > T_c$  (lower). Extracted from Ref. [32].

The analysis above is valid only for pure glue, i.e. with no dynamical quarks. However, we can still ask whether  $Z(3)$  is an approximate symmetry in QCD. On the lattice, in full QCD, one sees a remarkable variation of  $\ell$  around  $T_c$ , so that it plays the role of an approximate order parameter [33]. Notice, however, that  $Z(3)$  is broken at high, not low  $T$ , just the opposite of what is found in the analogous description of spin systems, such as Ising, Potts, etc [13]. The effective potential for the Polyakov loop is illustrated in Fig. 6.



#### 2.4 Adding quarks: chiral symmetry

In the limit of massless quarks, QCD is invariant under global chiral rotations  $U(N_f)_L \times U(N_f)_R$  of the quark fields. One can rewrite this symmetry in terms of vector ( $V = R + L$ ) and axial ( $A = R - L$ ) rotations

$$U(N_f)_L \times U(N_f)_R \sim U(N_f)_V \times U(N_f)_A . \quad (22)$$

As  $U(N) \sim SU(N) \times U(1)$ , one finds

$$U(N_f)_L \times U(N_f)_R \sim SU(N_f)_L \times SU(N_f)_R \times U(1)_V \times U(1)_A , \quad (23)$$

where we see the  $U(1)_V$  from quark number conservation and the  $U(1)_A$  broken by instantons.

In QCD, the remaining  $SU(N_f)_L \times SU(N_f)_R$  is explicitly broken by a nonzero mass term. Take, for simplicity,  $N_f = 2$ . Then,

$$\mathcal{L} = \frac{1}{4} F_{\mu\nu}^a F^{\mu\nu a} + \bar{\psi}_L \gamma^\mu D_\mu \psi_L + \bar{\psi}_R \gamma^\mu D_\mu \psi_R - m_u (\bar{u}_L u_R + \bar{u}_R u_L) - m_d (\bar{d}_L d_R + \bar{d}_R d_L) , \quad (24)$$

so that, for non-vanishing  $m_u = m_d$ , the only symmetry that remains is the vector isospin  $SU(2)_V$ . In the light quark sector of QCD, chiral symmetry is just approximate. Then, for massless QCD, one should find parity doublets in the vacuum, which is not confirmed in the hadronic spectrum. Thus, chiral symmetry must be broken in the vacuum by the presence of a quark chiral condensate, so that

$$SU(N_f)_L \times SU(N_f)_R \mapsto SU(N_f)_V , \quad (25)$$

and the broken generators allow for the existence of pions, kaons, etc.

Hence, for massless QCD, we can define an order parameter for the spontaneous breaking of chiral symmetry in the vacuum - the chiral condensate:

$$\langle 0 | \bar{\psi} \psi | 0 \rangle = \langle 0 | \bar{\psi}_L \psi_R | 0 \rangle + \langle 0 | \bar{\psi}_R \psi_L | 0 \rangle , \quad (26)$$

so that this vacuum expectation value couples together the  $L$  and  $R$  sectors, unless in the case it vanishes. For very high temperatures or densities (low  $\alpha_s$ ), one expects to restore chiral symmetry, melting the condensate that is a function of  $T$  and quark masses and plays the role of an order parameter for the chiral transition in QCD.

Again, the analysis above is valid only for massless quarks. However, we can still ask whether QCD is approximately chiral in the light quark sector. On the lattice (full massive QCD), one sees a remarkable variation of the chiral condensate around  $T_c$ , so that the condensate plays the role of an approximate order parameter [33].

In summary, there are two relevant phase transitions in QCD, associated with spontaneous symmetry breaking mechanisms for different symmetries of the action: (i) an approximate  $Z(N_c)$  symmetry and deconfinement, which is exact for pure gauge  $SU(N_c)$  with an order parameter given by the Polyakov loop; (ii) an approximate chiral symmetry and chiral transition, which is exact for massless quarks, with an order parameter given by the chiral condensate.

One can try to investigate these phase transitions by building effective models based on such symmetries of the QCD action. Then, the basic rules would be: (i) keeping all relevant symmetries of the action; (ii) trying to include in the effective action all terms allowed by the chosen symmetries; (iii) developing a mimic of QCD at low energy using a simpler field theory; (iv) providing, whenever possible, analytic results at least for estimates and qualitative behavior. Well-known examples are the linear sigma model, the Nambu-Jona-Lasinio model, Polyakov loop models and so on [24]. Although they represent just part of the story, combined with lattice QCD they may provide good insight.

### 3 A final comment

Instead of conclusions, just a final comment on a point we have already made in the discussion above. To make progress in understanding, or at least in collecting facts about, (de)confinement and chiral symmetry, we need it all: experiments and observations, lattice simulations, theory developments, effective models, and also combinations whenever possible. In that vein, it is absolutely crucial to have theorists and experimentalists working and discussing together.

### Acknowledgements

The work of ESF was financially supported by the Helmholtz International Center for FAIR within the framework of the LOEWE program (Landesoffensive zur Entwicklung Wissenschaftlich-Ökonomischer Exzellenz) launched by the State of Hesse.

### References

- [1] G. Altarelli, hep-ph/0204179.
- [2] K. Fukushima, J. Phys. G **39** (2012) 013101.
- [3] D. J. Gross and F. Wilczek, Phys. Rev. Lett. **30** (1973) 1343; H. D. Politzer, Phys. Rev. Lett. **30** (1973) 1346.
- [4] T. van Ritbergen, J. A. M. Vermaseren and S. A. Larin, Phys. Lett. B **400** (1997) 379.
- [5] G. 't Hooft, hep-th/0010225.
- [6] C.-N. Yang and R. L. Mills, Phys. Rev. **96** (1954) 191.
- [7] Clay Mathematics Institute, <http://www.claymath.org/millennium/Yang-Mills-Theory/>.
- [8] J. F. Donoghue, E. Golowich and B. R. Holstein, *Dynamics of the standard model*, Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol. **2** (1992) 1.
- [9] S. Pokorski, *Gauge Field Theories* (Cambridge Monographs On Mathematical Physics, 1987).
- [10] F. Wilczek, Phys. Today **53N8** (2000) 22.
- [11] S. Borsanyi, G. Endrodi, Z. Fodor, S. D. Katz and K. K. Szabo, JHEP **1207** (2012) 056.
- [12] E. S. Fraga, Y. Hatta, R. D. Pisarski and J. Schaffner-Bielich, nucl-th/0301062.
- [13] L. E. Reichl, *A Modern Course in Statistical Physics* (Wiley, 2009).
- [14] M. G. Alford, A. Schmitt, K. Rajagopal and T. SchLfer, Rev. Mod. Phys. **80** (2008) 1455.
- [15] Jan Klaers *et al.*, Nature **468** (2010) 545.
- [16] M. A. Stephanov, Prog. Theor. Phys. Suppl. **153** (2004) 139 [Int. J. Mod. Phys. A **20** (2005) 4387].
- [17] E. W. Kolb and M. S. Turner, *The Early Universe*, Front. Phys. **69** (1990) 1; S. Weinberg, *Cosmology* (Oxford University Press, 2008).
- [18] C. Y. Wong, *Introduction to High-Energy Heavy Ion Collisions* (World Scientific, Singapore, 1994). L. P. Csernai, *Introduction to Relativistic Heavy Ion Collisions* (John Wiley and Sons, Chichester, 1994). J. Harris and B. Müller, Ann. Rev. Nucl. Part. Sci. **46** (1996) 71.
- [19] L. McLerran and N. Samios, *T. D. Lee: Relativistic Heavy Ion Collisions and the Riken Brookhaven Center*, BNL-77850-2007-CP.
- [20] *RHIC – Relativistic Heavy Ion Collider*, <http://www.bnl.gov/rhic/>.
- [21] *LHC – Large Hadron Collider*, <http://home.web.cern.ch/about/experiments/>.
- [22] *FAIR – Facility for Antiproton and Ion Research*, <http://www.gsi.de/fair/>.
- [23] *NICA – Nuclotron-based Ion Collider Facility*, <http://nica.jinr.ru/>.
- [24] D. H. Rischke, Prog. Part. Nucl. Phys. **52** (2004) 197; K. Yagi, T. Hatsuda and Y. Miake, *Quark-Gluon Plasma: From Big Bang To Little Bang*, Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol. **23** (2005) 1.

- [25] S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (Wiley, 1972).
- [26] L. F. Palhares, E. S. Fraga and T. Kodama, *J. Phys.* **38** (2011) 085101.
- [27] T. Ullrich, B. Wyslouch and J. W. Harris, *Nucl. Phys. A* **904-905** (2013) pp. 1c.
- [28] N. K. Glendenning, *Compact Stars — Nuclear Physics, Particle Physics, and General Relativity* (Springer, New York, 2000).
- [29] J. I. Kapusta and C. Gale, *Finite-Temperature Field Theory: Principles and Applications* (Cambridge University Press, 2006); M. Le Bellac, *Thermal Field Theory* (Cambridge University Press, 1996).
- [30] O. Philipsen, *Prog. Part. Nucl. Phys.* **70** (2013) 55.
- [31] R. D. Pisarski, hep-ph/0203271.
- [32] A. J. Mizher, M. N. Chernodub and E. S. Fraga, *Phys. Rev. D* **82** (2010) 105016.
- [33] S. Borsanyi, S. Durr, Z. Fodor, C. Hoelbling, S. D. Katz, S. Krieg, D. Nogradi and K. K. Szabo *et al.*, *JHEP* **1208** (2012) 126.



# Ultra-High Energy Cosmic Rays

*M.T. Dova*

Instituto de Física La Plata, Universidad Nacional de La Plata and CONICET, Argentina

## Abstract

The origin of the ultra high energy cosmic rays (UHECR) with energies above  $E > 10^{17}$  eV, is still unknown. The discovery of their sources will reveal the engines of the most energetic astrophysical accelerators in the universe. This is a written version of a series of lectures devoted to UHECR at the 2013 CERN-Latin-American School of High-Energy Physics. We present an introduction to acceleration mechanisms of charged particles to the highest energies in astrophysical objects, their propagation from the sources to Earth, and the experimental techniques for their detection. We also discuss some of the relevant observational results from Telescope Array and Pierre Auger Observatory. These experiments deal with particle interactions at energies orders of magnitude higher than achieved in terrestrial accelerators.

## 1 Introduction

Extreme physical systems provide the best scenario to study the fundamental physical laws. In this direction the research on ultra high energy cosmic rays is a crucial element, contributing to progress in both astrophysics and particle physics. UHECR open a window to energy and kinematic regions previously unexplored in the study of fundamental interactions and continue to motivate current and future cosmic ray experiments. In this note we summarize a series of lectures given at the 7th CERN-Latin-American School of High-Energy Physics on ultra high energy cosmic rays, the highest-energy particles measured on Earth with energy  $E > 10^{17}$  eV.

UHECR are mainly protons and nuclei, accelerated in astrophysical objects. The requirements for these objects to be sources of UHECR are quite stringent, as in addition to be able to accelerate to extremely high energies, they should also have the luminosity that can account for the observed fluxes. UHECR must survive during acceleration, escape and propagation through the intergalactic space, losing energy in the interactions with the Infrared/optical (IR/O), Cosmic Microwave Background (CMB) or Radio Background photons. We begin with a brief introduction to cosmic rays. Then, we introduce basic concepts of acceleration mechanisms, and the main energy loss processes for UHECR during propagation. The opacity of the CMB to the propagation of these particles is a key issue in the search for the origin of UHECR, leading to a modification of the energy spectrum and a strong constraint on the proximity of UHECR sources. At this point we give a short description of the main experimental techniques for the detection of UHECR and discuss observational results of the cosmic ray spectrum. UHECR are also deflected in the intergalactic and galactic magnetic fields in the propagation volume, what limits the search for correlations of the arrival direction of UHECR with possible sources and distributions of astrophysical objects in our vicinity. Here we present studies of anisotropy at the highest energies. Next, we summarize the phenomenology of cosmic ray air showers, including the dominant electromagnetic processes driving the shower evolution. We also present the hadronic interaction models used to extrapolate results from collider data to ultrahigh energies. Finally, we describe the main observables sensitive to primary composition, the most challenging issue to understand the nature and origin of UHECR.

## 2 Cosmic Rays

In 1912, Victor Hess carried out a series of balloon flights taking an electroscope to measure the ionizing radiation as a function of altitude. He discovered that the ionization rate increased by at least a factor

of two at around 5 km above the Earth’s surface [1]. He received the Nobel prize in 1936 for the discovery of this “penetrating radiation” coming from space, later called cosmic rays. In 1938, Pierre Auger and his colleagues first reported the existence of extensive air showers (EAS), showers of secondary particles caused by the collision of primary high energy particles with air molecules. On the basis of his measurements, Auger concluded that he had observed showers with energies of  $10^{15}$  eV [2, 3]. The literature abounds in historical introductions to cosmic rays, we recommend the heart-warming notes by J. Cronin at the 30th International Cosmic Ray Conference [4]. See also the lectures notes presented in Refs. [5, 6].

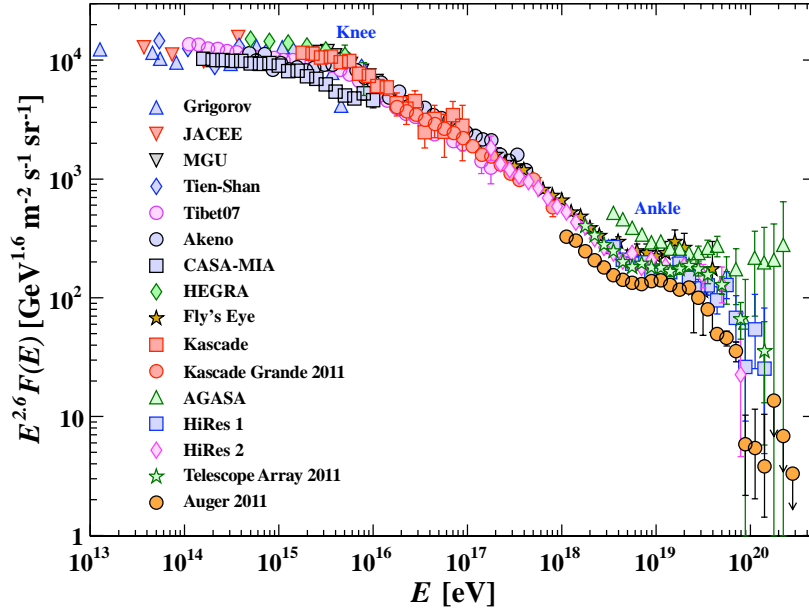
For primary energy above  $10^{11}$  eV, the observed cosmic ray flux can be described by a series of power laws with the flux falling about three orders of magnitude for each decade increase in energy. Figure 1 shows the “all-particle” spectrum. The differential energy spectrum has been multiplied by  $E^{2.6}$  in order to display the features of the steep spectrum that are otherwise difficult to discern [7]. A change of the spectral index ( $E^{-2.7}$  to  $E^{-3.0}$ ) at an energy of about  $10^{15}$  eV is known as the cosmic ray knee. This feature is generally believed to correspond to the steepening of the galactic proton spectrum, either because a change of the propagation regime or because of maximum limitations at the source, [8–10]. The same effect for heavier nuclei may cause the softer spectrum above the knee. In this context, subsequent steepenings of the spectrum are predicted at  $E_{max} \sim Z \times 10^{15}$  eV reaching  $\sim 8 \times 10^{16}$  eV for the iron group. The KASCADE-Grande collaboration provided the first observation of this sequence of changes [11]. Above several  $\sim 10^{18}$  eV the magnetic field in the vicinity of the Galaxy would not trap very effectively even the very heaviest nuclei, so the detected cosmic rays must be extragalactic [12]. The onset of an extragalactic contribution could be indicated by the so-called second knee, a further steepening of the spectrum at about  $10^{17.7}$  eV. The flattening around  $10^{18.5}$  eV is called the ankle of the spectrum. The simplest way of producing this feature is that of intersecting the steep galactic spectrum with a flatter extragalactic one. Under this assumption, several models have been developed. In the “ankle model” [13, 14], the transition appears at  $10^{18.5}$  eV. This model needs a new high energy galactic component between the iron knee and the onset of the extragalactic component. In the “dip model”, the ankle appears as an intrinsic part of the pair-production dip, a feature predicted in the spectrum of extragalactic protons that can be directly linked to the interaction of UHECR with the CMB [15–17]. In this model the transition from the galactic to the extragalactic component begins at the second knee and is completed at the beginning of the dip at  $E \sim 10^{18}$  eV. In “mix composition models” [18], the transition occurs at  $3 \times 10^{18}$  eV with mass composition changing from the galactic iron to extragalactic mixed composition of different nuclei. For a recent comprehensive review of the transition models see Ref. [19].

The Large Hadron Collider (LHC) will collide in 2015 protons at  $\sqrt{s} \simeq 14$  TeV. This impressive energy is still about a factor of 50 smaller than the centre-of-mass energy of the highest energy cosmic ray so far observed, assuming primary protons.

For cosmic ray energies above  $10^{15}$  eV, the flux becomes so low that direct detection of the primary using devices in or above the upper atmosphere is, for all practical purposes, impossible. Fortunately, in such cases the primary particle has enough energy to initiate a particle cascade in the atmosphere large enough that the products are detectable at ground. There are several techniques which can be employed in detecting these extensive air showers (EAS), ranging from sampling of particles in the cascade to measurements of fluorescence, Čerenkov or radio emissions produced by the shower.

### 3 Acceleration of cosmic rays

There are two types of mechanisms able to accelerate charged particles to reach ultrahigh energies and at the same time give a power law injection spectrum. One is the acceleration of particles directly to very high energy by an extended electric field [20], such as the case of unipolar inductors in relativistic magnetic rotators (e.g. neutron stars [21]) or black holes with magnetized disks that lose rotational energy in jets. They have the advantage of being fast, however, they suffer from the circumstance that



**Fig. 1:** All-particle spectrum of cosmic rays. From Ref. [7]

the acceleration occurs in astrophysical sites of very high energy density, where new opportunities for energy loss exist. In addition, they predict a hard injection spectrum that cannot be reconciled with the currently observed slope. In 1949, Fermi introduced a statistical acceleration mechanism [22]. In his publication, Fermi considered the scattering of cosmic particles on moving magnetized clouds which led to a fractional energy gain  $\xi = \langle \Delta E \rangle / E \propto \beta^2$  where  $\beta$  is the average velocity of the scattering centres in units of  $c$ . There is a net transfer of the macroscopic kinetic energy from the moving cloud to the particle, but the average energy gain is very small. Nowadays, this process is called “second order Fermi acceleration”. The first really successful theory of high energy cosmic ray acceleration was identified in [23] to be the Fermi acceleration in nonrelativistic shock waves in supernova remnants. The diffusion of cosmic rays in moving magnetized plasmas in the upstream and downstream of the shocks, force particles to repeatedly cross the shock front, hence gaining energy by numerous encounters, this results in  $\xi \propto \beta$ . When measured in the stationary upstream frame,  $\beta$  is the speed of the shocked fluid in units of  $c$ . This mechanism is known as “first order Fermi acceleration”. Shock waves for UHECR acceleration are Gamma Ray Bursts (GRB) shocks, jets and hot spots of Active Galactic Nuclei (AGN), and gravitational accretion shocks.

Following [24], we provide here a simple calculation to obtain the power law predictions from first order Fermi processes under the “test particle approximation”, in which the back-reaction of accelerated CRs on the shock properties is neglected. The energy  $E_n$  of a cosmic particle after  $n$  acceleration cycles is:

$$E_n = E_0(1 + \xi)^n \quad (1)$$

and the number of cycles to reach  $E$  results from Eq. (1)

$$n = \ln \left( \frac{E}{E_0} \right) / \ln(1 + \xi) \quad (2)$$

where  $E_0$  is the energy at injection into the acceleration site. If the escape probability  $P_{esc}$  per encounter

is constant, then the probability to stay in the acceleration region after  $n$  cycles is  $(1 - P_{esc})^n$ . The fraction of particles accelerated to energies  $> E$ , the integral spectrum, is:

$$N(> E) \propto \frac{(1 - P_{esc})^n}{P_{esc}} \propto \frac{1}{P_{esc}} \left( \frac{E}{E_0} \right)^{-\gamma} \quad (3)$$

with  $\gamma \propto P_{esc}/\xi$  for  $\xi \ll 1$  and  $P_{esc} \ll 1$ . Note that both first and second order Fermi acceleration produce a power law energy spectrum.

The escape probability from the acceleration site depends on the characteristic time for the acceleration cycle and the characteristic time for escape from the acceleration site. In the rest frame of the shock the conservation relations imply that the upstream velocity  $u_{up}$  is much higher than the downstream velocity  $u_{down}$ . The compression ratio  $r = u_{up}/u_{down} = n_{down}/n_{up}$  can be determined by requiring continuity of particle number, momentum, and energy across the shock. Here  $n_{up}$  ( $n_{down}$ ) is the particle density of the upstream (downstream) plasma. For an ideal gas the compression ratio can be related to the specific heat ratio and the Mach number of the shock. In the case of highly supersonic shocks,  $r = 4$  [25]. To determine the spectrum we need to calculate  $\gamma$ . For the case of shock acceleration,  $\xi = 4\beta/3 = 4(u_{up} - u_{down})/3$  and the escape probability can be obtained as the ratio of the loss flux, downstream away from the shock, and the crossing flux. Assuming the configuration of a large, plane shock the escape probability results as  $P_{esc} = 4u_{down}/c$ . Finally, we obtain the spectral index of the integral energy spectrum:

$$\gamma \propto P_{esc}/\xi \propto \frac{3}{u_{up}/u_{down} - 1} \propto 1 \quad (4)$$

This injection spectrum should be compared with the observed flux of cosmic rays,  $dN/dE \propto E^{-2}$ . The result is in good agreement although additional effects, like energy losses or an energy dependence of the escape probability, could have an important impact on the shape of the injection spectrum. For a comprehensive review of shock acceleration theory, see Ref. [25]. For a discussion about different acceleration mechanisms we recommend Ref. [26].

The requirements for astrophysical objects to be sources of UHECR are stringent. The Larmor radius of a particle with charge  $Ze$  increases with its energy  $E$  according to

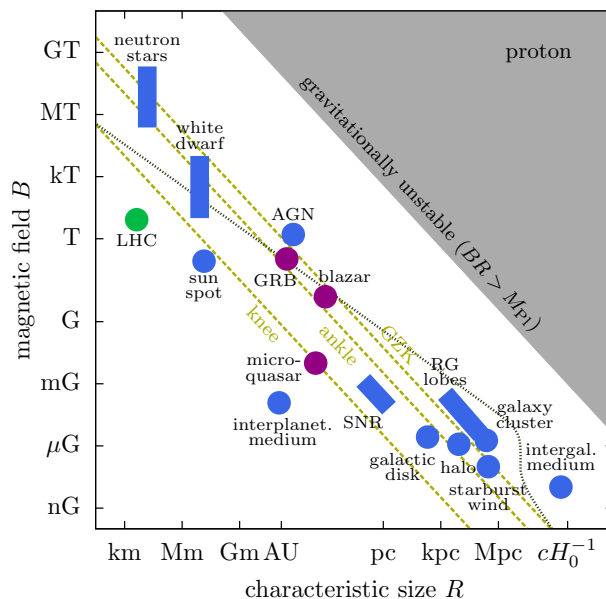
$$r_L = \frac{1.1}{Z} \left( \frac{E}{10^{18} \text{eV}} \right) \left( \frac{B}{\mu\text{G}} \right)^{-1} \text{ kpc}. \quad (5)$$

The search for UHECR extragalactic sources was motivated by the fact that  $r_L$  in the galactic magnetic field is much larger than the thickness of the galactic disk, hence, confinement in the galaxy is not maintained for UHECR. The famous Hillas criteria states that the Larmor radius of the accelerated particles cannot exceed the size of the source ( $R_{\text{source}}$ ), setting a natural limit in the particle's energy.

$$E_{\text{max}} \simeq Z \left( \frac{B}{\mu\text{G}} \right) \left( \frac{R_{\text{source}}}{\text{kpc}} \right) \times 10^{18} \text{ eV}. \quad (6)$$

This limitation in energy can be seen in the so-called Hillas plot [27] shown in Fig. 2 where candidate sources are placed in a plane of the characteristic magnetic field  $B$  versus their characteristic size  $R$ . For protons, the only sources for the UHECR that seem to be plausible are radio galaxy lobes and clusters of galaxies. Exceptions may occur for sources which move relativistically in the host-galaxy frame, in particular jets from AGN and GRB. In this case the maximal energy might be increased due to a Doppler boost by a factor  $\sim 30$  or  $\sim 1000$ , respectively. For a survey of cosmic ray sources shown in Fig. 2 and their signatures, see Refs. [26, 28]. An interesting point is that if acceleration takes place in GRB, one may expect a strong neutrino signature due to proton interactions with the radiative background [29]. Such a signature is now being probed by the Ice Cube experiment [30].





**Fig. 2:** The “Hillas plot” for various CR source candidates (blue). Also shown are jet-frame parameters for blazars, gamma-ray bursts, and microquasars (purple). The corresponding point for the LHC beam is also shown. The red dashed lines show the *lower limit* for accelerators of protons at the CR knee ( $\sim 10^{14.5}$  eV), CR ankle ( $\sim 10^{18.5}$  eV) and the GZK suppression ( $\sim 10^{19.6}$  eV). The dotted gray line is the *upper limit* from synchrotron losses and proton interactions in the cosmic photon background ( $R \gg 1$  Mpc). From Ref. [31].

## 4 Propagation of extragalactic cosmic rays

### 4.1 Energy losses of protons

There are three main energy loss processes for protons propagating over cosmological distances: Adiabatic energy losses due to the expansion of the universe,  $-dE/dt = H_0$ , pair production ( $p\gamma \rightarrow pe^+e^-$ ) and pion-production  $p\gamma \rightarrow \pi N$  on photons of the cosmic microwave background (CMB). Collisions with optical and infrared photons give a negligible contribution.

The fractional energy loss due to interactions with the cosmic background radiation at a redshift  $z = 0$  is determined by the integral of the nucleon energy loss per collision multiplied by the probability per unit time for a nucleon collision in an isotropic gas of photons [32]. For interactions with a blackbody field of temperature  $T$ , the photon density is that of a Planck spectrum, so the fractional energy loss is given by

$$-\frac{1}{E} \frac{dE}{dt} = -\frac{ckT}{2\pi^2\Gamma^2(c\hbar)^3} \sum_j \int_{\omega_{0j}}^{\infty} d\omega_r \sigma_j(\omega_r) y_j \omega_r \ln(1 - e^{-\omega_r/2\Gamma kT}), \quad (7)$$

where  $\omega_r$  is the photon energy in the rest frame of the nucleon, and  $y_j$  is the inelasticity, *i.e.* the average fraction of the energy lost by the photon to the nucleon in the laboratory frame for the  $j$ th reaction channel. The sum is carried out over all channels and  $d\omega$ ,  $\sigma_j(\omega_r)$  is the total cross section of the  $j$ th interaction channel,  $\Gamma$  is the usual Lorentz factor of the nucleon, and  $\omega_{0j}$  is the threshold energy for the  $j$ th reaction in the rest frame of the nucleon.

At energies  $E \ll m_e m_p/kT = 2.1 \times 10^{18}$  eV, the reaction ( $p\gamma \rightarrow pe^+e^-$ ) takes place on the photons from the high energy tail of the Planck distribution. The cross section of the reaction approximated by the threshold values is  $\sigma(\omega_r) = \frac{\pi}{12} \alpha r_0^2 \left(\frac{\omega_r}{m_e} - 2\right)^3$ ,  $\alpha$  is the fine structure constant and  $r_0$  is the classical radius of the electron [33]. The inelasticity at threshold results  $y = 2 \frac{m_e}{m_p}$ . The fractional

energy loss due to pair production is then,

$$-\frac{1}{E} \left( \frac{dE}{dt} \right) = \frac{16c}{\pi} \frac{m_e}{m_p} \alpha r_0^2 \left( \frac{kT}{hc} \right)^3 \left( \frac{\Gamma kT}{m_e} \right)^2 \exp\left(-\frac{m_e}{\Gamma kT}\right). \quad (8)$$

At higher energies ( $E > 10^{19}$  eV) the photopion reactions  $p\gamma \rightarrow p\pi^0$  and  $p\gamma \rightarrow \pi^+n$  on the tail of the Planck distribution give the main contribution to proton energy loss. The photons are seen blue-shifted by the cosmic rays in their rest frames and the reaction becomes possible. The cross sections of these reactions are well known. It strongly increase at the  $\Delta(1232)$  resonance, which decays into the one pion channels  $\pi^+n$  and  $\pi^0p$  at a photon energy in the proton rest frame of 145 MeV. At higher energies, heavier baryon resonances occur and the proton might reappear only after successive decays of resonances. The cross section in this region can be described by a sum of Breit-Wigner distributions over the main resonances produced in  $N\gamma$  collisions with  $\pi N$ ,  $\pi\pi N$  and  $K\Lambda$  ( $\Lambda \rightarrow N\pi$ ) final states [34]. For the cross section at high energies the fits from the CERN-HERA and COMPAS Groups to the high-energy  $p\gamma$  cross section [35] can be used. Assuming that reactions mediated by baryon resonances have spherically symmetric decay angular distributions, the average energy loss of the nucleon after  $n$  resonant collisions is given by

$$y_\pi(m_{R_0}) = 1 - \frac{1}{2^n} \prod_{i=1}^n \left( 1 + \frac{m_{R_i}^2 - m_M^2}{m_{R_{i-1}}^2} \right), \quad (9)$$

where  $m_{R_i}$  denotes the mass of the  $i^{\text{th}}$  resonant system of the decay chain,  $m_M$  the mass of the associated meson,  $m_{R_0} = \sqrt{s}$  is the total energy of the reaction in the c.m., and  $m_{R_n}$  the mass of the nucleon. It is well established from experiments that, at very high energies ( $\sqrt{s} > 3$  GeV), the incident nucleons lose one-half their energy via pion photoproduction independent of the number of pions produced (“leading particle effect”) [36].

A fit to Eq. (7) for the region  $\sqrt{s} < 2$  GeV with the exponential behavior derived from the values of cross section and fractional energy loss at threshold, gives [37]

$$-\frac{1}{E} \left( \frac{dE}{dt} \right)_\pi = A \exp[-B/E], \quad (10)$$

$$A = (3.66 \pm 0.08) \times 10^{-8} \text{ yr}^{-1}, \quad B = (2.87 \pm 0.03) \times 10^{11} \text{ GeV}. \quad (11)$$

The fractional energy loss at higher c.m. energies ( $\sqrt{s} \gtrsim 3$  GeV) is roughly a constant,

$$-\frac{1}{E} \left( \frac{dE}{dt} \right)_\pi = C = (2.42 \pm 0.03) \times 10^{-8} \text{ yr}^{-1}. \quad (12)$$

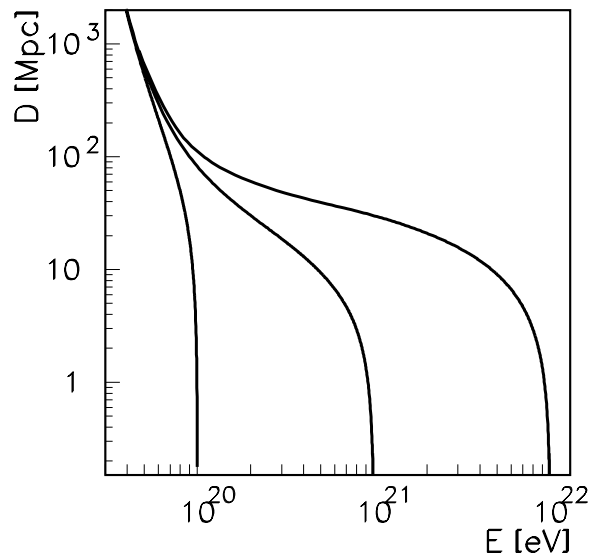
From the values determined for the fractional energy loss, it is straightforward to compute the energy degradation of UHECRs in terms of their flight time. This is given by,

$$At - \text{Ei}(B/E) + \text{Ei}(B/E_0) = 0, \quad \text{for } 10^{10} \text{ GeV} \lesssim E \lesssim 10^{12} \text{ GeV}, \quad (13)$$

and

$$E(t) = E_0 \exp[-Ct], \quad \text{for } E \gtrsim 10^{12} \text{ GeV}, \quad (14)$$

where Ei is the exponential integral. Figure 3 shows the proton energy degradation as a function of the mean propagation distance. Notice that, independent of the initial energy of the nucleon, the mean energy values approach  $10^{20}$  eV after a distance of  $\approx 100$  Mpc. This fact constrains the proximity to the Earth of the sources of UHECR with energies above  $5 \times 10^{19}$  eV.



**Fig. 3:** Energy attenuation length of protons in the intergalactic medium. For proton sources beyond  $\approx 100$  Mpc, the observed proton energy is  $< 10^{20}$  eV regardless its initial value. From Ref. [37].

## 4.2 Energy losses of nuclei

The relevant mechanisms for the energy loss of nuclei during propagation are: Compton interactions, pair production in the field of the nucleus, photodisintegration and hadron photoproduction. For nuclei of energy  $E > 10^{19}$  eV the dominant loss process is photodisintegration. In the nucleus rest-frame, pair production has a threshold at  $\sim 1$  MeV, photodisintegration is particularly important at the peak of the giant dipole resonance (15 to 25 MeV), and photomeson production has a threshold energy of  $\sim 145$  MeV. Compton interactions result in only a negligibly small energy loss for the nucleus [38].

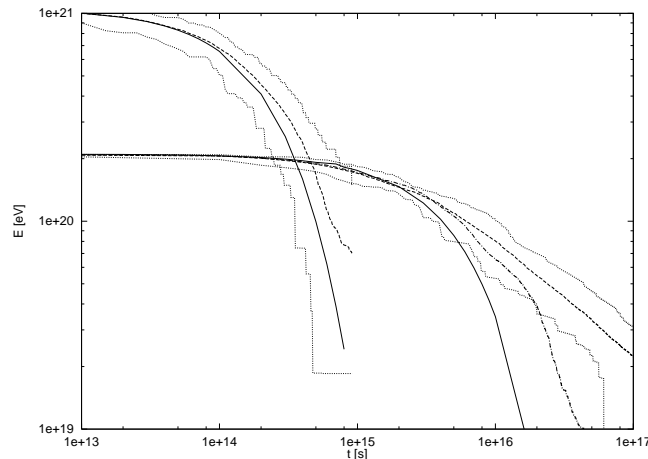
For a nucleus of mass  $A$  and charge  $Ze$ , the energy loss rate due to photopair production is  $Z^2/A$  times higher than for a proton of the same Lorentz factor [39], whereas the energy loss rate due to photomeson production remains roughly the same. The latter is true because the cross section for photomeson production by nuclei is proportional to the mass number  $A$  [40], while the inelasticity is proportional to  $1/A$ . However, it is photodisintegration rather than photopair and photomeson production that determines the energetics of ultrahigh energy cosmic nuclei. During this process some fragments of the nuclei are released, mostly single neutrons and protons. Experimental data of photonuclear interactions are consistent with a two-step process: photoabsorption by the nucleus to form a compound state, followed by a statistical decay process involving the emission of one or more nucleons.

The disintegration rate with production of  $i$  nucleons is given by [41]

$$R_{Ai} = \frac{1}{2\Gamma^2} \int_0^\infty dw \frac{n(w)}{w^2} \int_0^{2\Gamma w} dw_r w_r \sigma_{Ai}(w_r) \quad (15)$$

where  $n(w)$  is the density of photons with energy  $w$  in the system of reference in which the cosmic microwave background (CMB) is at 2.7 K and  $w_r$  is the energy of the photons in the rest frame of the nucleus. As usual,  $\Gamma$  is the Lorentz factor and  $\sigma_{Ai}$  is the cross section for the interaction.

Here, the soft photon background is taken as the sum of a 2.7 K Planckian spectrum that dominates at energies  $w \in (2.0 \times 10^{-6} \text{ eV}, 4 \times 10^{-3} \text{ eV})$ , and the infrared radiation as estimated in Ref. [42]. Parameterizations of the photodisintegration cross section for the different nuclear species are given in Ref. [38]. Summing over all possible channels for a given number of nucleons, one obtains the effective nucleon loss rate  $R = \sum_i i R_{Ai}$ . The effective nucleon loss rate for light elements, as well as for those in



**Fig. 4:** The energy of the surviving fragment ( $\Gamma_0 = 4 \times 10^9$ ,  $\Gamma_0 = 2 \times 10^{10}$ ) vs. propagation time obtained using Eq. (20) is indicated with a solid line. Also included is the energy attenuation length obtained from Monte Carlo simulations with (dashed) and without (dotted-dashed) pair creation production, for comparison. The region between the two dotted lines includes 95% of the simulations. This gives a clear idea of the range of values which can result from fluctuations from the average behaviour.

the carbon, silicon and iron groups can be scaled as in [38]

$$\left. \frac{dA}{dt} \right|_A \sim \left. \frac{dA}{dt} \right|_{\text{Fe}} \left( \frac{A}{56} \right) = R|_{\text{Fe}} \left( \frac{A}{56} \right), \quad (16)$$

with the photodisintegration rate parametrized by [43]

$$R_{56}(\Gamma) = 3.25 \times 10^{-6} \Gamma^{-0.643} \exp(-2.15 \times 10^{10}/\Gamma) \text{ s}^{-1} \quad (17)$$

for  $\Gamma \in [1.0 \times 10^9, 36.8 \times 10^9]$ , and

$$R_{56}(\Gamma) = 1.59 \times 10^{-12} \Gamma^{-0.0698} \text{ s}^{-1} \quad (18)$$

for  $\Gamma \in [3.68 \times 10^{10}, 10.0 \times 10^{10}]$ .

For photodisintegration, the averaged fractional energy loss results equal to the fractional loss in mass number of the nucleus, because the nucleon emission is isotropic in the rest frame of the nucleus. During the photodisintegration process the Lorentz factor of the nucleus is conserved, unlike the cases of pair production and photomeson production processes which involve the creation of new particles that carry off energy. The total fractional energy loss is then

$$-\frac{1}{E} \frac{dE}{dt} = \frac{1}{\Gamma} \frac{d\Gamma}{dt} + \frac{R}{A}. \quad (19)$$

For  $\omega_r \lesssim 145$  MeV the reduction in  $\Gamma$  comes from the nuclear energy loss due to pair production [44]. For  $\Gamma > 10^{10}$  the energy loss due to photopair production is negligible, and thus

$$E(t) \sim 938 A(t) \Gamma \text{ MeV} \sim E_0 e^{-R(\Gamma)|_{\text{Fe}} t/56}. \quad (20)$$

Figure 4 shows the energy of the heaviest surviving nuclear fragment as a function of the propagation time, for initial iron nuclei. The solid curves are obtained using Eq. (20), whereas the dashed and dotted-dashed curves are obtained by means of Monte Carlo simulations [45]. One can see that nuclei

with Lorentz factors above  $10^{10}$  cannot survive for more than 10 Mpc. For these distances, the approximation given in Eq. (20) always lies in the region which includes 95% of the Monte Carlo simulations. When the nucleus is emitted with a Lorentz factor  $\Gamma_0 < 5 \times 10^9$ , pair production losses start to be relevant, significantly reducing the value of  $\Gamma$  as the nucleus propagates distances of  $\mathcal{O}(100 \text{ Mpc})$ . The effect has a maximum for  $\Gamma_0 \approx 4 \times 10^9$  but becomes small again for  $\Gamma_0 \leq 10^9$ , for which appreciable effects only appear for cosmological distances ( $> 1000 \text{ Mpc}$ ), see for instance Ref. [45].

Note that Eq. (20) imposes a strong constraint on the location of nucleus-sources: less than 1% of iron nuclei (or any surviving fragment of their spallations) can survive more than  $3 \times 10^{14} \text{ s}$  with an energy  $> 10^{20.5} \text{ eV}$ . It is important to keep in mind that a light propagation distance of  $1.03 \times 10^{14} \text{ s}$  corresponds to 1 Mpc.

In recent years the interest in the propagation of UHECR nuclei has significantly grown. A complete review with a detailed list of references can be found in [46]. Most recent calculations of UHECR proton propagation use the Monte Carlo generator SOPHIA [47] for photomeson interaction of protons, based on available data and phenomenological models. For the case of nuclei propagation, existing propagation codes are CRPropa [48] and the complete nuclei propagation tool presented in Ref. [49].

## 5 Cosmic ray observations at the highest energies: Hybrid instruments

For primary cosmic ray energies above  $10^{14} \text{ eV}$ , the flux becomes so low that individual events cannot longer be detected directly. Fortunately, in such cases the primary particle has enough energy to initiate an extended air shower (EAS) in the atmosphere. Only the secondary particles are detected and used to infer the properties of the primary particle. There are several techniques which can be employed in detecting EAS.

The most commonly used detection method involves sampling the shower front at a given altitude using an array of sensors spread over a large area. The classical set up consists of an array of plastic scintillators, registering charged particles from the shower (also some converted photons). Another technique is to use water Čerenkov detectors (WCD), that allow the detection of the very numerous photons present in showers. They are deep compare with scintillators, so they have larger response to inclined showers. An initial estimate of the shower direction is obtained from the relative arrival times of signal at a minimum of 3 non-collinear detectors, treating the shower front as if it were planar. The density of particles falls off with the distance to the shower core and this can be parameterized by a lateral distribution function (LDF), which, of course, depends on the characteristics of the detectors used. The particle density at a large distance from the shower core is commonly used as an energy estimator. Muons in the EAS have higher energies than electromagnetic particles, which in addition suffer significant scattering and energy loss. Thus, the muonic component tends to arrive earlier and over a shorter period of time than the electromagnetic one. These signatures may also help to distinguish  $\mu$ 's from electrons and  $\gamma$ 's providing a useful tool to determine the primary composition.

Another highly successful air shower detection method involves measurement of the longitudinal development of the cascade by sensing the fluorescence light produced via interactions of the charged particles in the atmosphere. As an extensive air shower develops, it dissipates much of its energy by exciting and ionizing air molecules along its path. Excited nitrogen molecules fluoresce producing radiation in the 300 - 400 nm ultraviolet range, to which the atmosphere is quite transparent. Under favourable atmospheric conditions EAS can be detected at distances as large as 20 km, though observations can only be made on clear moonless nights, yielding a duty cycle of about 10%. The shower development appears as a rapidly moving spot of light whose angular motion depends on both the distance and the orientation of the shower axis. The fluorescence technique provides the most effective way to measure the energy of the primary particle. The amount of fluorescence light emitted is proportional to the number of charged particles in the showers allowing a direct measurement of the longitudinal development of the EAS in the atmosphere. For this, the sky is viewed by many segmented eyes using photomultipliers. From the

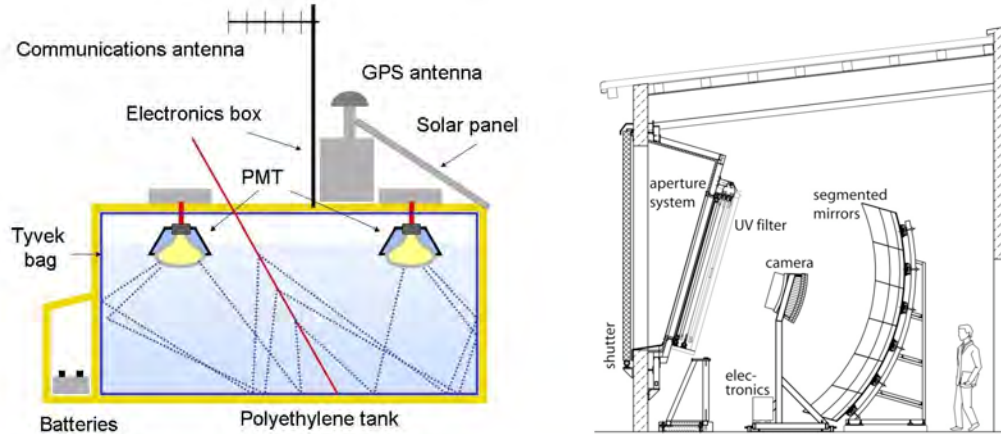
measured shower profile the position of the shower maximum  $X_{max}$ , which is sensitive to primary composition, can be obtained. The energy in the electromagnetic component is calculated by integrating the measured shower profile, after corrections for atmospheric attenuation of the fluorescence light and contamination of the signal by Čerenkov light. Finally, to derive the total energy of the shower, an estimate of the missing energy carried to the ground by neutrinos and high energy muons must be made based on assumptions about the primary mass and the appropriate hadronic interaction models.

In this note we focus on the two high energy cosmic ray experiments currently operating: the Pierre Auger Observatory [50] and the Telescope Array (TA) [51]. The Pierre Auger Observatory, the largest UHECR experiment in the world, is located in Malargüe, Argentina ( $35^{\circ}12'S$ ,  $69^{\circ}12'W$ ). It has an accumulated exposure of about  $30000 \text{ km}^2 \text{ sr yr}$ . The Telescope Array located in Millard County, Utah, USA ( $39.3^{\circ}N$ ,  $112.9^{\circ}W$ ), due to a later start and its more than 4 times smaller area, has collected about 10 times less events. Both the Pierre Auger Observatory and TA are hybrid detectors employing two complementary detection techniques for the ground-based measurement of air showers induced by UHECR: a surface detector array (SD) and a fluorescence detector (FD).

The ground array of the Pierre Auger Observatory consists of 1600 stations spaced by 1.5 km covering an area of  $3000 \text{ km}^2$ . Each detector is a cylindrical, opaque tank of  $10 \text{ m}^2$  and a water depth of 1.2 m, where particles produce light by Čerenkov radiation. The filtered water is contained in an internal coating which diffusely reflects the light collected by three photomultipliers (PMT) installed on the top. The large diameter PMTs ( $\approx 20 \text{ cm}$ ) hemispherical photomultiplier are mounted facing down and look at the water through sealed polyethylene windows that are integral part of the internal liner. Due to the size of the array the stations have to work in an autonomous way. Thus the stations operate on battery-backed solar power and communicate with a central station by using wireless LAN radio links. The time information is obtained from the Global Positioning Satellite (GPS) system. This array is fully efficient at energies above  $E > 3 \times 10^{18} \text{ eV}$ . Additional detectors with 750 m spacing have been nested within the 1500 m array to cover an area of  $25 \text{ km}^2$  with full efficiency above  $E > 3 \times 10^{17} \text{ eV}$ . The SD is sensitive to electromagnetic and muonic secondary particles of air showers and has a duty cycle of almost 100%. The surface array is overlooked by 27 optical telescopes grouped in 5 buildings on the periphery of the array [52]. The field of view of each telescope is  $30^{\circ}$  in azimuth, and  $1.5^{\circ}$  to  $30^{\circ}$  in elevation, except for three of them, for which the elevation is between  $30^{\circ}$  and  $60^{\circ}$  (HEAT telescopes [53]). Light is focused with a spherical mirror of  $13 \text{ m}^2$  on a camera of 440 hexagonal PMTs. The FD can only operate during dark nights, which limits its duty cycle to 13%. Stable data taking with the SD started in January 2004 and the Observatory has been running with its full configuration since 2008.

In Figure 5 (left panel) we present a schematic description of a water Čerenkov detector installed at the Pierre Auger Observatory. Mounted on top of the tank are the solar panel, electronic enclosure, mast, radio antenna and GPS antenna for absolute and relative timing. A battery is contained in a box attached to the tank. The main components of a fluorescence eye are shown on the right panel of Figure 5: a large spherical mirror with a radius of curvature of 3.4 m, a pixel camera in the focal surface and a diaphragm with an entrance glass window. This filter allows reduction of night background with respect to the fluorescence signal and also serves to protect the equipment from dust.

The TA surface array consists of 507 detector units deployed in a square grid with 1.2 km spacing to cover a total area of approximately  $700 \text{ km}^2$ . Each unit consists of a plastic scintillation counter of  $3 \text{ m}^2$  surface and 1.2 cm thickness, with 2 layers of plastic scintillators viewed by PMT at each end. The entire system is powered by a solar panel and battery. The communication is done with WLAN modem. The SD array is fully efficient for cosmic rays with energies greater than  $10^{18.8} \text{ eV}$  [54]. Three FD stations are placed around the SD array, with a total of 38 telescopes. Each telescope is comprised of a cluster of photo-tubes and a reflecting mirror of 3.3 m diameter. A PMT camera consisting of  $16 \times 16$  PMTs is set at a distance of 3000 mm from the mirror. The field of view of each PMT is approximately  $1^{\circ}$  and that of the FD station is from  $3^{\circ}$  to  $33^{\circ}$  in elevation and  $108^{\circ}$  in azimuth. See Ref. [51] for details of the TA detectors.



**Fig. 5:** Left: A typical surface detector of the Auger Observatory. Right: A fluorescence telescope. See the text for the description of the components.

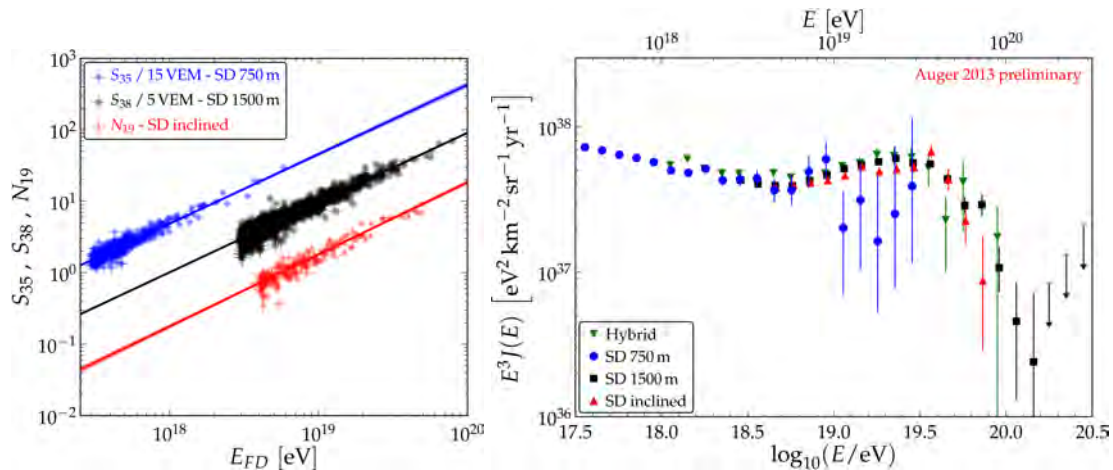
## 6 Flux measurements

Surface arrays, with its near 100% duty cycle, give the larger data sample used to obtain the energy spectrum. The comparison of the shower energy, measured using fluorescence, with the SD energy parameter for a subset of hybrid events is used to calibrate the energy scale for the array.

The first step towards the flux measurement with the SD array is the reconstruction of arrival direction and core position of air showers. Then, a stable parameter from the SD which correlates with the primary energy is reconstructed. This parameter is the signal at an optimal distances to the shower core at which the spread in the signal size is minimum [55]. In the following we distinguish between *vertical events* ( $\theta < 60^\circ$ ) and *inclined events* ( $62^\circ \leq \theta < 80^\circ$ ). For the case of Auger, the optimal distance is 1000 m for the main array and 450 m for the “infill”, while for TA is 800 m. For *vertical events* the signals at the optimal distance obtained from a LDF fit, have to be corrected for their zenith angle dependence due to air shower attenuation in the atmosphere. This is done in Auger with a Constant Intensity Cut (CIC) method [56]. The equivalent signal at median zenith angle of  $38^\circ$  ( $35^\circ$ ) is then used to infer the energy for the 1500 m (750 m) array [57, 58]. Events that have independently triggered the SD array and FD telescopes are used for the energy calibration of SD data [59]. The correlation between the different energy estimators and the energy obtained from the FD is shown in Figure 6 (left panel) superimposed with the calibration functions resulting from maximum-likelihood fits. For the case of TA, the energy is estimated by using a look-up table in  $S(800)$  and zenith angle determined from an exhaustive Monte Carlo simulation. The uncertainty in energy scale of the Monte Carlo simulation of an SD is large, and possible biases associated with the modelling of hadronic interactions are difficult to determine. Therefore, the SD energy scale is corrected to the TA FD using hybrid events. The observed differences between the FD and SD events are well described by a simple proportionality relationship, where the SD energy scale is 27% higher than the FD [60].

Water Čerenkov detectors from the Pierre Auger Observatory SD, have larger response to inclined showers. These EAS are characterized by the dominance of secondary muons at ground, as the electromagnetic component is largely absorbed in the large atmospheric depth traversed by the shower [61]. The reconstruction is based on the estimation of the relative muon content  $N_{19}$  with respect to a simulated proton shower with energy  $10 \times 10^{19}$  eV [62].  $N_{19}$  is used to infer the primary energy for inclined events, as shown in the left panel of Figure 6.

The energy spectra obtained from the three SD datasets are shown in the right panel of Figure 6. To characterize the spectral features, the Auger collaboration describes the data with a power law below



**Fig. 6:** Left: The correlation between the different energy estimators S38, S35 and N19 (see text) and the energy determined by FD. Right: Energy spectra, corrected for energy resolution, derived from SD and from hybrid data. From Ref. [57].

the ankle  $J(E) \propto E^{-\gamma_1}$  and a power law with smooth suppression above:

$$J(E; E > E_a) \propto E^{-\gamma_2} \left[ 1 + \exp \left( \frac{\log_{10} E - \log_{10} E_{1/2}}{\log_{10} W_c} \right) \right]^{-1}.$$

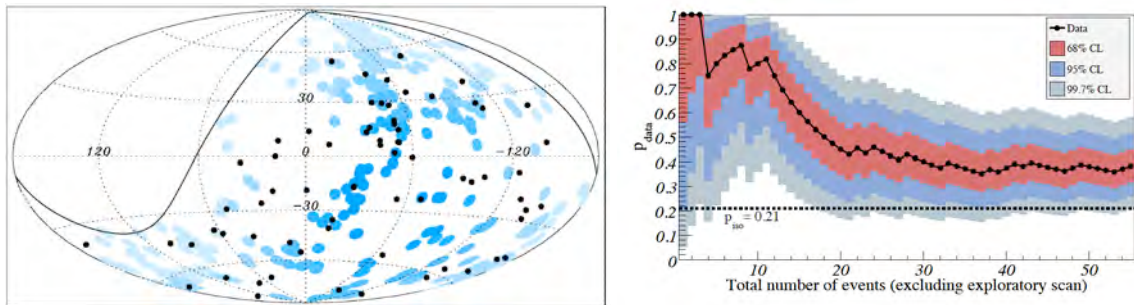
$\gamma_1$ ,  $\gamma_2$  are the spectral indices below/above the ankle at  $E_a$ .  $E_{1/2}$  is the energy at which the flux has dropped to half of its peak value before the suppression, the steepness of which is described with  $\log_{10} W_c$ . The data in Figure 6 clearly exhibit the ankle at  $10^{18.7} \text{eV}$  and a flux suppression above  $10^{19.6} \text{eV}$ . The Pierre Auger Observatory has confirmed the GZK feature of the spectrum with a significance greater than  $20 \sigma$  obtained by comparison to a power law extrapolation. This observation seems to indicate that acceleration in extragalactic sources can explain the high energy CR spectrum, ending the need for exotic alternatives designed to avoid the flux suppression. However, the possibility that this feature in the spectrum is due to the maximum energy of acceleration at the sources is not easily dismissed.

We present here only the energy spectrum from the Pierre Auger Observatory, details of the corresponding spectrum obtained by the Telescope Array collaboration are presented in Ref. [63]. As discussed in Ref. [64], it is found that the energy spectra determined by these experiments are consistent in normalization and shape after energy scaling factors are applied. Those scaling factors are within systematic uncertainties in the energy scale quoted by the experiments.

## 7 Correlation with astrophysical objects

Since the UHECR are charged particles, they not only lose energy in the interaction with background photons, but also they are deflected by galactic and extragalactic magnetic fields. The galactic magnetic field (GMF) can be modelled as the sum of a regular (large scale fluctuations) and a turbulent (smaller scale fluctuations) components. The directions on the sky in which cosmic rays are deflected strongly depend on the GMF model, however, averaged quantities such as the average UHECR deflection angle are much less model dependent [65]. Extragalactic magnetic fields are expected to be stronger in the large scale structure of the Universe and significantly weaker in voids. UHECR deflections in such fields are poorly constrained ranging from negligible to more than ten degrees, even for 100 EeV protons (See Ref. [26] and references therein). Attempts to detect anisotropies at ultrahigh energies are based on the





**Fig. 7:** Left: The 69 arrival directions of cosmic rays with energy  $E > 55$  EeV detected by the Pierre Auger Observatory up to December 2009 are plotted as black dots in an Aitoff-Hammer projection of the sky in galactic coordinates. The solid line represents the field of view of the Southern Observatory for zenith angles smaller than  $60^\circ$ . Blue circles of radius  $3.1^\circ$  are centred at the positions of the 318 AGN in the VCV catalogue that lie within 75 Mpc and that are within the field of view of the Observatory. Darker blue indicates larger relative exposure. The exposure-weighted fraction of the sky covered by the blue circles is 21%. Right: Fraction of events correlating with AGN as a function of the cumulative number of events, starting after the exploratory data. The expected correlating fraction for isotropic cosmic rays is shown by the dotted line. From Ref. [68]

selection of events with the largest magnetic rigidity to study whether they can be correlated with the direction of possible sources or distributions of astrophysical objects in our vicinity (less than 100 Mpc).

The most recent discussion of anisotropies in the sky distribution of ultrahigh energy events began when the Pierre Auger Observatory reported a correlation of its highest energy events with AGN [66] in the 12th Veron-Cetty & Veron (VCV) catalogue [67]. To calculate a meaningful statistical significance in such an analysis, it is important to define the search procedure *a priori* in order to ensure it is not inadvertently devised especially to suit the particular data set after having studied it. With the aim of avoiding accidental bias on the number of trials performed in selecting the cuts, the Auger anisotropy analysis scheme followed a pre-defined process. First an exploratory data sample was employed for comparison with various source catalogues and for tests of various cut choices. The results of this exploratory period were then used to design prescriptions to be applied to subsequently gathered data. The first 14 events were used for an exploratory scan and the correlation was most significant for AGN for energy threshold  $5.5 \times 10^{19}$  eV with redshifts  $z < 0.018$  (distances  $< 75$  Mpc) and within  $3.1^\circ$  separation angles. The subsequent 13 events established a 99% confidence level for rejecting the hypothesis of isotropic cosmic ray flux. The reported fraction of correlation events was  $69^{+11}_{-13}\%$ . An analysis with data up to the end of 2009 (69 events in total, as seen in the left panel of Figure 7) indicated that the correlation level decreased to  $38^{+7}_{-6}\%$  [68]. In the right panel of Figure 7 we show the most likely value of the fraction of the correlated events with objects in the VCV catalogue as a function of the total number of time-ordered events (the events used in the exploratory scan are excluded). The  $1\sigma$  and  $2\sigma$  uncertainties in this value are indicated. The current estimate of the fraction of correlating cosmic rays is  $33 \pm 5\%$  (28 events correlating from a total of 84 events) with 21% expected under the isotropic hypothesis [69].

The Telescope Array Collaboration has also searched for correlation with AGN in the VCV catalogue [70, 71]. The TA exposure is peaked in the Northern hemisphere so the AGN visible to TA are not the same as the ones visible to Auger, though there is some overlap. When the distribution of nearby AGN is taken into account, and assuming equal AGN luminosities in UHECR, the correlating fraction would be 40%.

A complete report on the current status for anisotropy searches can be found in [72]. The report includes, in the region around  $10^{18}$  eV, constraints from measuring the first harmonic modulation in the right ascension distribution of arrival directions, and search for point-like sources that would be indicative

of a flux of neutrons (see also Ref. [73]); at higher energies, searches for clustering in arrival directions, and correlations with nearby extragalactic objects (see also Ref. [74]) or the large scale structure of the Universe.

## 8 Mass composition estimate: the biggest challenge

A determination of primary composition is invaluable in revealing the origin of cosmic rays as this information would provide important bounds on sources and on possible production and acceleration mechanisms. In addition, a proper interpretation of anisotropy information requires knowledge of the primary mass due to the influence on propagation of the galactic and intergalactic magnetic fields. A detailed analysis of composition data from various experiments has been presented in Ref. [75]. We first present a brief description of the general signatures of the EAS (See Ref. [76] for a summary of the phenomenology of these giant air showers). After that, we introduce the shower observables sensitive to primary species.

### 8.1 Signatures of Extensive Air Showers

The evolution of an extensive air shower is dominated by electromagnetic processes. The interaction of a baryonic cosmic ray with an air nucleus high in the atmosphere leads to a cascade of secondary mesons and nucleons. The first few generations of charged pions interact again, producing a hadronic core, which continues to feed the electromagnetic and muonic components of the showers. Up to about 50 km above sea level, the density of atmospheric target nucleons is  $n \sim 10^{20} \text{ cm}^{-3}$ , and so even for relatively low energies, say  $E_{\pi^\pm} \approx 1 \text{ TeV}$ , the probability of decay before interaction falls below 10%. Ultimately, the electromagnetic cascade dissipates around 90% of the primary particle's energy, and hence the total number of electromagnetic particles is very nearly proportional to the shower energy.

By the time a vertically incident  $10^{20} \text{ eV}$  proton shower reaches the ground, there are about  $10^{11}$  secondaries with energy above 90 keV in the annular region extending 8 m to 8 km from the shower core. Of these, 99% are photons, electrons, and positrons, with a typical ratio of  $\gamma$  to  $e^+e^-$  of 9 to 1. Their mean energy is around 10 MeV and they transport 85% of the total energy at ground level. Of course, photon-induced showers are even more dominated by the electromagnetic channel, as the only significant muon generation mechanism in this case is the decay of charged pions and kaons produced in  $\gamma$ -air interactions [77].

It is worth mentioning that these figures dramatically change for the case of very inclined showers. For a primary zenith angle,  $\theta > 70^\circ$ , the electromagnetic component becomes attenuated exponentially with atmospheric depth, being almost completely absorbed at ground level. As a result, most of the energy at ground level from an inclined shower is carried by muons.

In contrast to hadronic collisions, the electromagnetic interactions of shower particles can be calculated very accurately from quantum electrodynamics. Electromagnetic interactions are thus not a major source of systematic errors in shower simulations. The first comprehensive treatment of electromagnetic showers was elaborated by Rossi and Greissen [78]. This treatment was recently cast in a more pedagogical form by Gaisser [24], which we summarize in the subsequent paragraphs.

The generation of the electromagnetic component is driven by electron bremsstrahlung and pair production [79]. Eventually the average energy per particle drops below a critical energy,  $\epsilon_0$ , at which point ionization takes over from bremsstrahlung and pair production as the dominant energy loss mechanism. The  $e^\pm$  energy loss rate due to bremsstrahlung radiation is nearly proportional to their energy, whereas the ionization loss rate varies only logarithmically with the  $e^\pm$  energy. Throughout this note we take the critical energy to be that at which the ionization loss per radiation length is equal to the electron energy, yielding  $\epsilon_0 = 710 \text{ MeV}/(Z_{\text{eff}} + 0.92) \sim 86 \text{ MeV}$  [80]. The changeover from radiation losses to ionization losses depopulates the shower. One can thus categorize the shower development in three phases: the growth phase, in which all the particles have energy  $> \epsilon_0$ ; the shower maximum,  $X_{\text{max}}$ ; and

the shower tail, where the particles only lose energy, get absorbed or decay.

Most of the general features of an electromagnetic cascade can be understood in terms of the toy model due to Heitler [81]. In this model, the shower is imagined to develop exclusively via bremsstrahlung and pair production, each of which results in the conversion of one particle into two. These physical processes are characterized by an interaction length  $X_0$ . One can thus imagine the shower as a particle tree with branches that bifurcate every  $X_0$ , until they fall below a critical energy,  $\epsilon_0$ , at which point energy loss processes dominate. Up to  $\epsilon_0$ , the number of particles grows geometrically, so that after  $n = X/X_0$  branchings, the total number of particles in the shower is  $N \approx 2^n$ . At the depth of shower maximum  $X_{\max}$ , all particles are at the critical energy,  $\epsilon_0$ , and the energy of the primary particle,  $E_0$ , is split among all the  $N_{\max} = E_0/\epsilon_0$  particles. Putting this together, we get:

$$X_{\max} \approx X_0 \frac{\ln(E_0/\epsilon_0)}{\ln 2}. \quad (21)$$

Even baryon-induced showers are dominated by electromagnetic processes, so this toy model is still enlightening for such cases. In particular, for proton showers, Eq. (21) tells us that the  $X_{\max}$  scales logarithmically with primary energy, while  $N_{\max}$  scales linearly. Moreover, to extend this discussion to heavy nuclei, we can apply the superposition principle as a reasonable first approximation. In this approximation, we pretend that the nucleus comprises unbound nucleons, such that the point of first interaction of one nucleon is independent of all the others. Specifically, a shower produced by a nucleus with energy  $E_A$  and mass  $A$  is modelled by a collection of  $A$  proton showers, each with  $A^{-1}$  of the nucleus energy. Modifying Eq. (21) accordingly one easily obtains  $X_{\max} \propto \ln(E_0/A)$ .

Changes in the mean mass composition of the cosmic ray flux as a function of energy will manifest as changes in the mean values of  $X_{\max}$ . This change of  $X_{\max}$  with energy<sup>1</sup> is commonly known as the elongation rate theorem [82]:

$$D_e = \frac{\delta X_{\max}}{\delta \ln E}. \quad (22)$$

For purely electromagnetic showers,  $X_{\max}(E) \approx X_0 \ln(E/\epsilon_0)$  and then the elongation rate is  $D_e \approx X_0$ . For proton primaries, the multiplicity rises with energy, and thus the resulting elongation rate becomes smaller. This can be understood by noting that, on average, the first interaction is determined by the proton mean free path in the atmosphere,  $\lambda_N$ . In this first interaction the incoming proton splits into  $\langle n(E) \rangle$  secondary particles, each carrying an average energy  $E/\langle n(E) \rangle$ . Assuming that  $X_{\max}(E)$  depends logarithmically on energy, as we found with the Heitler model described above, it follows that,

$$X_{\max}(E) = \lambda_N + X_0 \ln[E/\langle n(E) \rangle]. \quad (23)$$

If we assume a multiplicity dependence  $\langle n(E) \rangle \approx n_0 E^\Delta$ , then the elongation rate becomes,

$$\frac{\delta X_{\max}}{\delta \ln E} = X_0 \left[ 1 - \frac{\delta \ln \langle n(E) \rangle}{\delta \ln E} \right] + \frac{\delta \lambda_N}{\delta \ln E} \quad (24)$$

which corresponds to the form given in [83],

$$D_e = X_0 \left[ 1 - \frac{\delta \ln \langle n(E) \rangle}{\delta \ln E} + \frac{\lambda_N}{X_0} \frac{\delta \ln \lambda_N}{\delta \ln E} \right] = X_0 (1 - B). \quad (25)$$

Using the superposition model and assuming that

$$B \equiv \Delta - \frac{\lambda_N}{X_0} \frac{\delta \ln \lambda_N}{\delta \ln E} \quad (26)$$

<sup>1</sup>The elongation rate is commonly reported per decade of energy,  $D_{10} = \partial \langle X_{\max} \rangle / \partial \log E$ , where  $D_{10} = 2.3 D_e$ .

is not changing with energy, one obtains for mixed primary composition [83]

$$D_e = X_0 (1 - B) \left[ 1 - \frac{\partial(\ln A)}{\partial \ln E} \right]. \quad (27)$$

Thus, the elongation rate provides a measurement of the change of the mean logarithmic mass with energy.

In Ref. [84], a precise calculation of a hadronic shower evolution has been presented assuming that hadronic interactions produce exclusively pions. The first interaction diverts 1/3 of the available energy ( $E_0/3$ ) into the EM component via the  $\pi^0$ 's, while the remaining 2/3 continue as hadrons. Using  $pp$  data [85], we parametrized the charged particle production in the first interaction as  $N_{\pi^\pm} = 41.2(E_0/1 \text{ PeV})^{1/5}$ . The depth of shower maximum is thus the same as for an electromagnetic shower of energy  $E_0/(3N_{\pi^\pm})$ , giving for a proton initiated shower:

$$\begin{aligned} X_{\text{max}}^p &= X_0 + X_{\text{EM}} \ln[E_0/(6N_{\pi^\pm}\epsilon_0)] \\ &= (470 + 58 \log_{10}[E_0/1 \text{ PeV}]) \text{ g/cm}^2. \end{aligned} \quad (28)$$

For protons the elongation rate results  $\approx 58 \text{ g/cm}^2$  per decade of energy, in good agreement with calculations that model the shower development using the best estimates of the relevant features of the hadronic interactions. Muons are produced from the pion decay when they reach the critical energy ( $\xi_c^\pi$ ) after  $n_c$  generations. Introducing  $\beta = \ln(2N_\pi)/\ln(3N_\pi)$ , the total number of muons is:

$$N_\mu = (E_0/\xi_c^\pi)^\beta. \quad (29)$$

For  $N_\pi = 5$ ,  $\beta = 0.85$ . Unlike the electron number, the muon multiplicity does not grow linearly with the primary energy, but at a slower rate. The precise value of  $\beta$  depends on the average pion multiplicity used. It also depends on the inelasticity of the hadronic interactions. The critical pion energy  $\xi_c^\pi \approx 20 \text{ GeV}$  in a shower generated by 1 PeV proton.

Using the superposition model, we obtain for a nucleus of mass  $A$ .

$$N_\mu^A = A \left[ \frac{(E_0/A)}{\xi_c^\pi} \right]^\beta. \quad (30)$$

From the discussion above, it follows that the depth of shower maximum and the number of muons depend on the mass of the primary particle: iron initiated showers develop faster in the atmosphere, having smaller  $X_{\text{max}}$  than proton initiated shower, while larger number of muons are expected for heavier nuclei.

While the Heitler model is very useful for imparting a first intuition regarding global shower properties, the details of shower evolution are far too complex to be fully described by a simple analytical model. Full Monte Carlo simulation of interaction and transport of each individual particle is required for precise modelling of the shower development. At present two Monte Carlo packages are available to simulate EAS: CORSIKA (COsmic Ray SIMulation for KAscade) [86] and AIRES (AIR shower Extended Simulation) [87]. Both programs provide fully 4-dimensional simulations of the air showers initiated by protons, photons, and nuclei. A comparative study using these codes can be found in Ref. [88]. Different hadronic interaction models are used in these event generators, such as SIBYLL [89], QGSJET [90] and EPOS [91,92]. The LHC data, particularly those measured in the extreme forward region of the collisions, is of great importance to the physics of EAS. As an example, EPOS has been modified to reproduce in detail LHC data from various experiments [93].

## 8.2 Measurement of mass sensitive observables

In this section, we discuss how baryonic species may, to some extent, be distinguished by the signatures they produce in the atmosphere. The estimate of primary masses is the most challenging task in high energy cosmic ray physics as such measurements rely on comparisons of data to models. EAS simulations

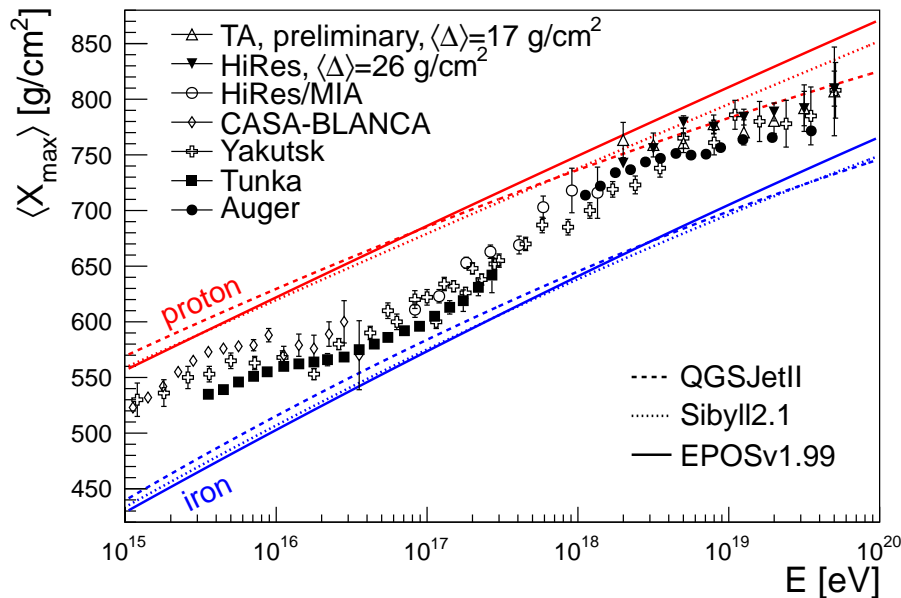
are subject to uncertainties mostly because hadronic interaction models need to be extrapolated at energy ranges several order of magnitude higher than those accessible to current particle accelerators. In what follows, we consider both surface array and fluorescence detector observables.

The main purpose of fluorescence detectors is to measure the properties of the longitudinal development. The shower longitudinal profile is usually parameterized with a function, such as the Gaisser-Hillas function [94] used by the Pierre Auger Observatory. Using this parameterization, fluorescence detectors can measure  $X_{\max}$  with a statistical precision typically around  $30 \text{ g/cm}^2$ . The speed of shower development is the clearest indicator of the primary composition. It was shown in Sec. 8 using the superposition model that there is a difference between the depth of maximum in proton and iron induced showers. In fact, nucleus-induced showers develop faster, having  $X_{\max}$  higher in the atmosphere. From Monte Carlo simulations, one finds that the difference between the average  $X_{\max}$  for protons and iron nuclei is about  $90 - 100 \text{ g/cm}^2$ . However, because of shower-to-shower fluctuations, it is not possible to obtain meaningful composition estimates from  $X_{\max}$  on a shower-by-shower basis, though one can derive composition information from the magnitude of the fluctuations themselves. For protons, the depth of first interaction fluctuates more than it does for iron, and consequently the fluctuations of  $X_{\max}$  are larger for protons as well. In Figure 8 the  $\langle X_{\max} \rangle$  measurements of  $\langle X_{\max} \rangle$  with non-imaging Cherenkov detectors (Tunka [96], Yakutsk [97], CASA-BLANCA [98]) and fluorescence detectors (HiRes/MIA [99], HiRes [100], Auger [101] and TA [102]) compared to air shower simulations using several hadronic interaction models are presented. The conclusion of the detailed study in Ref. [75] indicates that, around the region of the ankle of the cosmic ray spectrum, the measurements are compatible within their quoted systematic uncertainties and the  $\langle X_{\max} \rangle$  is close to the prediction for air showers initiated by a predominantly light composition. However, at higher energies, the experimental uncertainties are still too large to draw conclusions from the data. In addition, the systematic differences between different type of measurements are very sensitive to the particular interaction model used for the interpretation.

The electromagnetic component of an EAS suffers more scattering and energy loss than the muonic component and consequently, muons tend to arrive earlier and over a shorter period of time. This means that parameters characterizing the time structure of the EAS, as measured by surface arrays, will be correlated with  $X_{\max}$  and hence with primary mass. An early study of the shower signal observed in water Čerenkov detectors arrays [103] established the utility of a shower property known as risetime in estimating the primary composition. Specifically, the risetime,  $t_{1/2}$ , is defined as the time for the signal to rise from 10% to 50% of the full signal.

In ground array experiments the analysis is usually performed by projecting the signals registered by the detectors into the shower plane (perpendicular to the shower axis) and thus, neglecting the further shower evolution of the late regions. As a consequence, for inclined showers, the circular symmetry in the signals of surface detectors is broken. This results in a dependence of the signal features on the azimuth angle in the shower plane [104, 105]. A detailed study based on Monte Carlo simulations [106], showed that for showers arriving with zenith angle  $\theta > 30^\circ$ , this is mainly due to the attenuation of the electromagnetic component of the shower as it crosses additional atmosphere to reach a late detector. For a given primary energy  $E$ , the risetime asymmetry in water Čerenkov detectors array, as in the Pierre Auger Observatory, depends on zenith angle  $\theta$  of the primary cosmic ray in such a way that its behaviour versus  $\sec \theta$  is reminiscent of the longitudinal development of the shower. In Ref. [106], it was shown that the zenith angle at which the risetime asymmetry becomes maximum,  $\Theta_{max}$ , is correlated with the shower development and hence with the primary species.

Using the time information of the signals recorded by the water Čerenkov detectors, it is also possible to obtain information about the longitudinal development of the hadronic component of extensive air showers and the first interaction point in an indirect way. In particular, a method was developed to reconstruct the Muon Production Depth (MPD), the distance to the production of the muon measured parallel to the shower axis, using the signals of detectors far from the core [107]. The MPD technique allows one to convert the time distribution of the signal recorded by the SD detectors into muon produc-



**Fig. 8:** Measurements of  $\langle X_{\max} \rangle$  with non-imaging Cherenkov detectors (Tunka [96], Yakutsk [97], CASA-BLANCA [98]) and fluorescence detectors (HiRes/MIA [99], HiRes [100], Auger [101] and TA [102]) compared to air shower simulations using hadronic interaction models. HiRes and TA data have been corrected for detector effects as indicated by the  $\langle \Delta \rangle$  values, to allow comparison with the unbiased measurement from Auger. This picture is taken from Ref. [75].

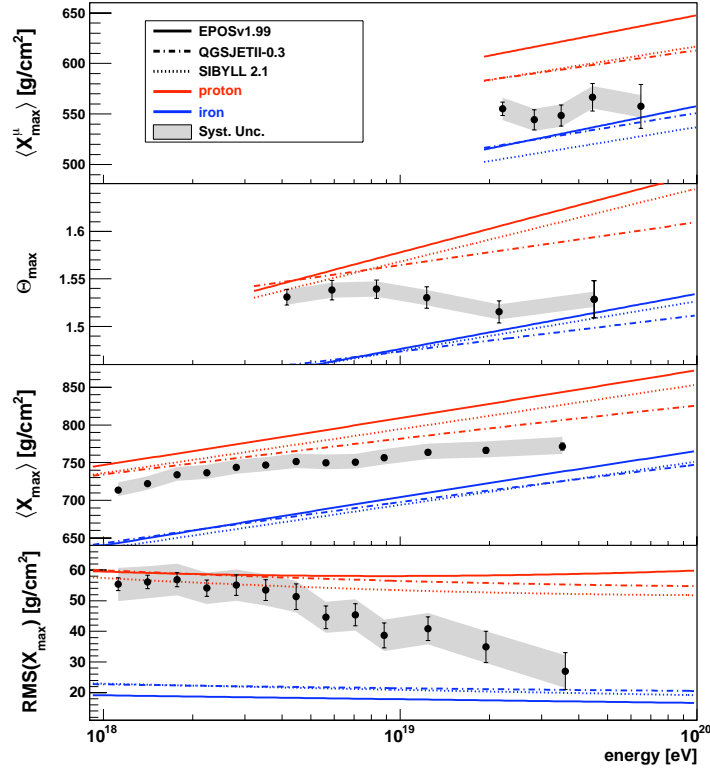
tion distances using an approximate relation between production distance, transverse distance and time delay with respect the shower front plane. From the MPDs a new observable can be defined,  $X_{\max}^{\mu}$ , as the depth along the shower axis where the number of produced muons reaches a maximum, which is sensitive to primary mass.

The evolution of  $X_{\max}^{\mu}$ ,  $\Theta_{\max}$ ,  $\langle X_{\max} \rangle$  and  $\text{RMS}(X_{\max})$  with energy, as measured by the Pierre Auger Observatory with data up to 2010 [108], is presented in Figure 9. For a very complete discussion of these results see Ref. [109]. It is worth noting that these analyses come from completely independent techniques that have different sources of systematic uncertainties. Concerning the RMS, a variety of compositions can give rise to large values of the RMS, because the width of the  $X_{\max}$  is influenced by both, the shower-to-shower fluctuations of individual components and their relative displacement in terms of  $\langle X_{\max} \rangle$ . These measurements from Auger may be interpreted as a transition to a heavier composition that may be caused by a Peters-cycle [110] in extragalactic sources similar to what has been observed at around the knee [75, 109].

Updated studies of  $X_{\max}^{\mu}$ ,  $\langle X_{\max} \rangle$  and  $\text{RMS}(X_{\max})$  from the Pierre Auger Observatory can be found in Ref. [111]. The most recent results on  $\langle X_{\max} \rangle$  measurements from the TA experiment were presented in Refs. [112, 113].

## Acknowledgments

I would like to thank the organizers of the 2013 CERN-Latin-American School of HEP for the excellent and stimulating school. I am indebted to Jim Cronin who introduced me to the fascinating world of cosmic rays, he has been an inspiration to me. I also would like to thank Luis Anchordoqui, Luis Epele and John Swain for the many years of fruitful discussions on the phenomenology of EAS and propagation of UHECR *en route* to us from their sources. I am grateful to Hernan Wahlberg, Paul Sommers, Michael



**Fig. 9:** From top to bottom,  $\langle X_{\max}^{\mu} \rangle$ ,  $\Theta_{\max}$ ,  $\langle X_{\max} \rangle$  and  $\text{RMS}(X_{\max})$  as a function of Energy compared with air shower simulations using different hadronic interaction models. The error bars correspond to the statistical uncertainty, the grey areas correspond to the systematic uncertainty [108]. Updated studies of  $X_{\max}^{\mu}$ ,  $\langle X_{\max} \rangle$  and  $\text{RMS}(X_{\max})$  can be found in Ref. [111].

Unger, Alan Watson, Analisa Mariazzi, Diego García-Pinto, Fernando Arqueros, Tom Paul and all my colleagues from the Pierre Auger Observatory for lively and enlightening discussions.

## References

- [1] V. F. Hess, Phys. Z. 13, 1804 (1912).
- [2] P. Auger, R. Maze, T. Grivet-Meyer, Comptes Rendus 206, 1721 (1938).
- [3] P. Auger, P. Ehrenfest, R. Maze, J. Daudin, Robley, and A. Freon, Rev. Mod. Phys. 11, 288 (1939).
- [4] James W. Cronin, Proceedings of the 30th International Cosmic Ray Conference, Universidad Nacional Autónoma de México, Mexico, Vol. 6, 3-19 (2009).
- [5] M. Kachelriess, Lecture Notes on High Energy Cosmic Rays, 2008, arXiv:0801.4376 [astro-ph].
- [6] L. Anchordoqui, Ultrahigh Energy Cosmic Rays: Facts, Myths, and Legends, 2011, arXiv:1104.0509 [hep-ph].
- [7] J. Beringer *et al.* [Particle Data Group], Phys. Rev. D86, 010001 (2012).
- [8] M. Aglietta *et al.* [EAS-TOP Collaboration], Astropart. Phys. 21, 583 (2004).
- [9] T. Antoni *et al.* [KASCADE Collaboration], Astropart. Phys. 24, 1 (2005).
- [10] W. D. Apel *et al.* [KASCADE Collaboration], Astropart. Phys. 31, 86 (2009).
- [11] W. D. Apel *et al.* [KASCADE-Grande Collaboration], Phys.Rev.Lett.107.171104 (2011)
- [12] A.M. Hillas, arXiv:0607109[astro-ph].
- [13] D. De Marco and T. Stanev, Phys. Rev. D72, 081301 (2005).

- [14] E. Waxman Nucl. Phys. B (Proc. Suppl) 87, 345 (2000).
- [15] V. S. Berezinsky, and S. I. Grigor'eva, Astron. Astrophys. 199, 1 (1988).
- [16] V. Berezinsky, A. Z. Gazizov, and S. I. Grigor'eva, Phys. ReV.D 74, 043005 (2006).
- [17] R. Aloisio, V. Berezinsky, P. Blasi, A. Gazizov, S. Grigor'eva, and B. Hnatyk, Astropart. Phys. 27, 76 (2007).
- [18] D. Allard *et al.* Astron. Astrophys. 443, L29 (2005); D. Allard, E. Parizot, and A. V. Olinto, Astropart. Phys. 27, 61 (2007); D. Allard *et al.* J. Phys. G 34, 359 (2007); D. Allard *et al.* JCAP 0810:033, (2008); C. De Donato, and G. A. Medina Tanco, Astropart. Phys. 32, 253 (2009).
- [19] R. Aloisio, V. Berezinsky and A. Gazizov. Astropart.Phys. 39-40, 129-143 (2012).
- [20] A. M. Hillas, Ann. Rev. Astron. Astrophys. 22, 425 (1984).
- [21] P. Blasi, R.I. Epstein, A. V. Olinto, ApJ Letters, 533, L123 (2000).
- [22] E. Fermi, Phys. Rev. 75, 1169 (1949).
- [23] W. I. Axford, E. Leer, and G. Skadron, International Cosmic Ray Conference, Vol. 11, 132-137 (1977); A. R. Bell, MNRAS 182, 147 (1978); R. D. Blandford and J. P. Ostriker, ApJ Letters, 221, L29 (1978).
- [24] T. K. Gaisser, *Cosmic Rays and Particle Physics*, (Cambridge University Press, 1990).
- [25] R. Blandford, D. Eichler, Phys. Rept. 154, 1-75 (1987).
- [26] K. Kotera, A.V. Olinto, Ann.Rev.Astron.Astrophys. 49, 119-153 (2011).
- [27] A. M. Hillas, Ann. Rev. Astron. Astrophys. 22, 425 (1984).
- [28] M. Lemoine, J.Phys.Conf.Ser. 409, 012007 (2013).
- [29] Waxman E., Bahcall J. Phys. Rev. Lett. 78 2292 (1997).
- [30] Abbasi R. *et al.* [The Ice Cube Collaboration], Nature 484, 351 (2012).
- [31] M. Ahlers, L. A. Anchordoqui, J. K. Becker, T. K. Gaisser, F. Halzen, D. Hooper, S. R. Klein. P. Mészáros, S. Razzaque, and S. Sarkar, FERMILAB-FN-0847-A, YITP-SB-10-01.
- [32] F. W. Stecker, Phys. Rev. Lett. **21**, 1016 (1968).
- [33] V. S. Berezinsky and S. I. Grigor'eva, Astron. Astrophys. 199, 1 (1988).
- [34] R. M. Barnett *et al.* [Particle Data Group], Phys. Rev. D 54, 1 (1996).
- [35] L. Montanet *et al.* [Particle Data Group], Phys. Rev. D 50, 1173 (1994). See p. 1335.
- [36] I. Golyak, Mod. Phys. Lett. A **7**, 2401 (1992).
- [37] L. A. Anchordoqui, M. T. Dova, L. N. Epele and J. D. Swain, Phys. Rev. D **55**, 7356 (1997).
- [38] J. L. Puget, F. W. Stecker and J. H. Bredekamp, Astrophys. J. **205**, 638 (1976).
- [39] M. J. Chodorowski, A. A. Zdziarski, and M. Sikora, Astrophys. J. **400**, 181 (1992).
- [40] S. Michalowski, D. Andrews, J. Eickmeyer, T. Gentile, N. Mistry, R. Talman and K. Ueno, Phys. Rev. Lett. **39**, 737 (1977).
- [41] F. W. Stecker, Phys. Rev. **180**, 1264 (1969).
- [42] M. A. Malkan and F. W. Stecker, Astrophys. J. **496** 13(1998).
- [43] L. A. Anchordoqui, M. T. Dova, L. N. Epele and J. D. Swain, Phys. Rev. D **57**, 7103 (1998).
- [44] F. W. Stecker and M. H. Salamon, Astrophys. J. **512**, 521 (1992).
- [45] L. N. Epele and E. Roulet, JHEP **9810**, 009 (1998).
- [46] Allard, D. Astropart. Phys 39-40, 33-43 (2012).
- [47] Mucke A., Engel R., Rachen J. P., Protheroe R. J., and Stanev T., Comp. Phys. Com., **124**, 290 (2000).
- [48] Armengaud, E., Sigl, G., and Miniati, F., Phys. Rev. D, **73**, 083008 (2006).
- [49] D. Allard, M. Ave., N. Busca *et al.* JCAP, **9**, 5 (2006).
- [50] J. Abraham *et al.* [Pierre Auger Collaboration], Nucl. Instrum. Meth. A **523**, 50 (2004).



- [51] T Abu-Zayyad *et al.* [TA Collaboration], Nucl. Instr. Meth. A689, 87 (2012).
- [52] J. Abraham *et al.* [Pierre Auger Collaboration], Nucl. Instrum. Meth.A 620, 227 (2010).
- [53] T. Hermann-Josef Mathes [Pierre Auger Collaboration], Proc. 32nd International Cosmic Ray Conference (ICRC 11), Beijing, China 3, 149 (2011).
- [54] D. Ivanov, B.T. Stokes *et al.* Proc. 32nd International Cosmic Ray Conference, 1297 (2011).
- [55] A.M. Hillas, Acta Physica Academiae Scientiarum Hungaricae 26, 355 (1970).
- [56] J. Hersil *et al.* Phys. Rev. Lett. 6, 22 (1961).
- [57] A. Schulz [Pierre Auger Collaboration], to appear in Proc. 33rd International Cosmic Ray
- [58] J. Abraham *et al.* [Pierre Auger Collaboration], Phys. Rev. Lett. 101, 061101(2008). R. Pesce [Pierre Auger Collaboration], Proc. 32nd International Cosmic Ray Conference (ICRC 11), Beijing, China (2011).
- [59] Conference (ICRC 13), Brasil (2013), arXiv:1307.5059[astro-ph].
- [60] T. Abu-Zayyad *et al.* ApJ768, L1 (2013).
- [61] I. Valino [Pierre Auger Collaboration], Proc. 31st International Cosmic Ray Conference (ICRC 09), Łódź, Poland (2009).
- [62] G. Rodriguez [Pierre Auger Collaboration], UHECR Symposium CERN (2012), EPJ Web Conf.53: 07003 (2013); I. Valino [Pierre Auger Collaboration] to appear in Proc. 33rd International Cosmic Ray Conference (ICRC 13), Brasil (2013).
- [63] O.E. Kalashev, E. Kido and the Telescope Array Collaboration, to appear in Proc. 33rd International Cosmic Ray Conference (ICRC 13), Brasil (2013).
- [64] B. Dawson, *et al.* [Pierre Auger, Telescope Array and Yakutsk Collaborations] Working Group report at UHECR Symposium CERN (2012), EPJ Web Conf.53: 01005 (2013).
- [65] G. Giacinti, M. Kachelriess, D. Semikoz and G. Sigl, G $\S$ nter. EPJ Web Conf. 53: 06004 (2013).
- [66] J. Abraham *et al.* [Pierre Auger Collaboration], Science 318 (5852), 938 (2007); J. Abraham *et al.* [Pierre Auger Collaboration], Astropart. Phys. 29, 188 (2008) [Erratum-ibid. 30, 45 (2008)].
- [67] M.-P. Veron-Cetty & P. Veron, Astron. Astrophys. 455, 773 (2006).
- [68] P. Abreu *et al.* [Pierre Auger Collaboration], Astropart. Phys. 34, 314 (2010)
- [69] K.-H. Kampert [Pierre Auger Collaboration], Proc. 32nd International Cosmic Ray Conference (ICRC211), Beijing (2011).
- [70] Abu-Zayyad, *et al.* Astrophys.J.757, 26 (2012).
- [71] M. Fukushima *et al.* [TA Collaboration], to appear in Proc. 33rd International Cosmic Ray Conference, Rio de Janeiro, Brazil (2013).
- [72] O. Deligny *et al.* [Pierre Auger, Yakutsk and Telescope Array Collaborations], Working Group report at UHECR Symposium CERN (2012), EPJ Web Conf.53:01008 (2013).
- [73] A. AAb *et al.* [Pierre Auger Collaboration], ApJ. 760, 148 (2012).
- [74] A. AAb *et al.* [Pierre Auger Collaboration], JCAP 05, 009 (2013)
- [75] K-H. Kampert and M. Unger, Astropart.Phys. 35, 660 (2012).
- [76] L. Anchordoqui, M. T. Dova, A. G. Mariazzi, T. McCauley, T. C. Paul, S. Reucroft and J. Swain, Annals Phys. 314, 145 (2004).
- [77] T. J. L. McComb, R. J. Protheroe and K. E. Turver, J. Phys. G 5, 1613 (1979).
- [78] B. Rossi and K. Greisen, Rev. Mod. Phys. 13, 240 (1941).
- [79] H. Bethe and W. Heitler, Proc. Roy. Soc. Lond. A 146, 83 (1934).
- [80] B. Rossi, *High Energy Particles* (Prentice-Hall, Inc., Englewood Cliffs, NY, 1952).
- [81] W. Heitler. *The Quantum Theory of Radiation*, 2nd. Edition, (Oxford Univ. Press, London, 1944).
- [82] J. Linsley. Proc. 15th International Cosmic Ray Conference, Plovdiv 12, 89 (1977).

- [83] J. Linsley and A. A. Watson, *Phys. Rev. Lett.* 46, 459 (1981).
- [84] J. Matthews, *Astropart. Phys.* 22, 387 (2005).
- [85] C. AMSLER *et al.* [Particle Data Group], *Phys. Lett. B* 667, 1 (2008).
- [86] D. Heck, G. Schatz, T. Thouw, J. Knapp and J. N. Capdevielle, FZKA-6019 (1998).
- [87] S. J. Sciutto, arXiv:9911331[astro-ph].
- [88] J. Knapp, D. Heck, S. J. Sciutto, M. T. Dova and M. Risse, *Astropart. Phys.* 19, 77 (2003).
- [89] R. S. Fletcher, T. K. Gaisser, P. Lipari and T. Stanev, *Phys. Rev. D* 50, 5710 (1994).
- [90] N. N. Kalmykov, S. S. Ostapchenko and A. I. Pavlov, *Nucl. Phys. Proc. Suppl.* 52B, 17 (1997).
- [91] K. Werner, F.-M. Liu, and T. Pierog, *Phys. Rev. C* 74, 044902 (2006).
- [92] T. Pierog and K. Werner, *Nucl. Phys. Proc. Suppl. B* 196,102 (2009).
- [93] T. Pierog, Iu Karpenko, J.M. Katzy, E. Yatsenko and K. Werner, arXiv:1306.0121[hep-ph].
- [94] T. K. Gaisser and A. M. Hillas. *Proc. of 15th International Cosmic Ray Conference, Plovdiv* 8, 353 (1977).
- [95] E. Barcikowski *et al.* [HiRes, Pierre Auger, Telescope Array and Yakutsk Collaborations] *Proc. of Int. Symposium on Future Directions in UHECR Physics (UHECR2012), CERN, Switzerland, EPJ Web Conf.* 53:01006 (2013).
- [96] N. Budnev *et al.*, [Tunka Collaboration] *Nucl. Phys. (Proc. Suppl.)* 190, 247(2009); V. Prosin *et al.* [Tunka Coll.], *Proc. 32nd International Cosmic Ray Conference, Beijing, China, vol. 1*, 197 (2011).
- [97] S. Knurenko, A. Sabourov [Yakutsk Collaboration], in: *Proceedings XVI ISVHECRI, 2010*; S. Knurenko, A. Sabourov, [Yakutsk Collaboration], *Nucl. Phys. B (Proc. Suppl.)* 212, 241 (2011).
- [98] J. Fowler *et al.* [CASA-BLANCA Collaboration], *Astropart. Phys.* 15, 49 (2001).
- [99] T. Abu-Zayyad *et al.* [HiRes/MIA Collaborations], *Astrophys. J.* 557, 686 (2001).
- [100] R. Abbasi *et al.* [HiRes collaboration], *Phys. Rev. Lett.* 104, 161101 (2010).
- [101] P. Facal *et al.* [Pierre Auger Collaboration] *Proc. 32nd International Cosmic Ray Conference, Beijing, China, vol. 2*, 105 (2011).
- [102] C. Jui *et al.* [TA Collaboration], *Proceedings APS DPF Meeting*, arXiv:1110.0133[astro-ph].
- [103] A. A. Watson and J. G. Wilson, *J. Phys. A* 7, 1199 (1974).
- [104] M. T. Dova [Pierre Auger Collaboration], *Proc. 28th International Cosmic Ray Conference, Tsukuba*, 369 (2003), arXiv:0308399[astro-ph].
- [105] M. T. Dova, L. N. Epele and A. G. Mariazzi, *Astropart. Phys.* 18, 351 (2003).
- [106] M. T. Dova, M. E. Manceñido, A. G. Mariazzi, H. Wahlberg, F. Arqueros and D. García-Pinto, *Astroparticle Phys.* 31, 312 (2009).
- [107] L. Cazon, R.A. Vazquez and E. Zas, *Astropart. Phys.* 23: 393-409 (2005); L. Cazon, R. Conceicao, M. Pimenta, E. Santos, *Astropart. Phys.* 36, 211 (2012).
- [108] D. García-Pinto [Pierre Auger Collaboration], *Proc. 32nd International Cosmic Ray Conference, Beijing, China* (2011).
- [109] M. Unger [Pierre Auger Coll.], *Proc. of Int. Symposium on Future Directions in UHECR Physics (UHECR2012), CERN, Switzerland, EPJ Web Conf.*, 53:04009 (2013).
- [110] B. Peters, *Nuovo Cimento* 22, 800 (1961).
- [111] A. Aab [Pierre Auger Collaboration], to appear in *Proc. 33rd International Cosmic Ray Conference, Rio de Janeiro, Brazil* (2013), arXiv:1307.5059[astro-ph].
- [112] W. Hanlon [TA Collaboration], to appear in *Proc. 33rd International Cosmic Ray Conference, Rio de Janeiro, Brazil* (2013).
- [113] Y. Tameda [TA Collaboration], to appear in *Proc. 33rd International Cosmic Ray Conference, Rio de Janeiro, Brazil* (2013).

## LHC Results Highlights

*O. González*  
CIEMAT, Spain

### Abstract

The good performance of the LHC provided enough data at 7 TeV and 8 TeV to allow the experiments to perform very competitive measurements and to expand the knowledge about the fundamental interaction far beyond that from previous colliders. This report summarizes the highlights of the results obtained with these data samples by the four large experiments, covering all the topics of the physics program and focusing on those exploiting the possibilities of the LHC.

### 1 Introduction

The standard model (SM) [1–3] of particles and interactions is currently the most successful theory describing the Universe at the smallest distances, or equivalently, highest energies. Such task is performed with the use of three families of fermions and a number of bosons associated to the interactions as given by the  $SU(3)_C \times SU(2)_L \times U(1)_Y$  symmetry group. Since in Nature the  $SU(2)_L \times U(1)_Y$  is not an exact symmetry, we require an additional field, the so-called Higgs field, which spontaneously breaks the symmetry according to the BEH mechanism [4], giving rise to the weak and electromagnetic interactions as they are observed at lower energies. In addition this field is responsible to give mass to the fermions.

Although successful, the SM does not appear to be complete since several experimental evidences are not included in the model. In this group, it should be remarked that gravitational effects are not described, neither are all the related effects, such as Dark Matter or Dark Energy. In addition the current structure of the SM does not include enough CP violation to justify the observed matter-antimatter imbalance in the Universe. Finally the neutrinos in the model are assumed to be massless, something that currently is experimentally discarded after the measurements of neutrino mixing.

In addition to the missing parts in the SM there are several points in which the model is not completely satisfactory, concretely related to theoretical aspects of it. Several issues are always mentioned in this context, but they are summarized in three main issues: the need of fine-tuning to understand the low scale of the electroweak symmetry breaking and other parameters (the hierarchy problem), the lack of understanding on why there are three families with double-nature sets (i.e. quark and leptons) and the lack of apparent relation between the different interactions (i.e. the origin of the observed values for couplings, including fermionic masses). In practice, the SM has clear limitations since it misses too many explanations about why things are as they are and it requires too many parameters to actually describe things as they are.

The proposed solution to both the experimentally-motivated limitations and the theoretical dissatisfaction is to add more interactions or particles which complete the model. In such scenario, the SM would become a low-energy approximation, or visible part, of a larger theory. By increasing the energy in our studies we gain access to the additional particles and effects, which are usually referred to as “new physics” or “physics beyond the SM” (BSM). These effects that are not explained by the SM will provide additional information about the limitations of the SM, opening the correct doors towards a more accurate description of our Universe.

With this motivation we are led to the design of a powerful hadronic collider which maximizes the reach in sensitivity to the possible BSM physics. This is achieved by maximizing the available energy, which would provide the possibility to produce more massive particles, and the number of collisions per time unit (luminosity), which increase the yield for the produced particles and effects. This is exactly

the motivation for the Large Hadron Collider (LHC) [5] located at CERN, near Geneva (Switzerland), which is recognized as “the discovery machine” for physics beyond the SM providing a large amount of energy per collision and a large amount of collisions.

In the following sections we will describe the LHC and the related experiments and report on the main results for the different part of the program, designed to take advantage of all the possibilities given by such powerful machines.

## 2 The LHC and the experiments

The LHC is the most energetic and most challenging collider up to date. It is designed to collide protons or heavy ions at a maximum energy of 14 TeV of energy and very high collision rates. Technical limitations has prevented it to reach its design parameters, and the collected datasets contains collisions at 7 or 8 TeV of total center-of-mass energies. In any case this represents more than 3 times more energy than the previously most energetic collider (The Tevatron at Fermilab, USA). This allows to reach energy scales that were not accessible before, both for particle and heavy-ion physics.

But the LHC is not just about large energy: it also provides the largest collision rate ever reached, allowing to collect sizable data samples in record time. To quantify the amount of data, the previously mentioned concept of luminosity is used. The integrated luminosity relates the number of a type of events in a sample and the cross section for that type of event. Experimentally, this allows to compute the luminosity (“calibrate” the size of the sample) using a very well known process and count the number of events from it, and so  $L = N/\sigma$  where  $L$  is the luminosity,  $N$  the number of events and  $\sigma$  the cross section of the process. Once the sample luminosity is known, the value is used to measure cross sections of processes of interest, as  $\sigma = L/N$ . Finally, knowing the cross section of a process, one estimates the number of expected events from that process in the sample with  $N = L \cdot \sigma$ . These are the basic tools to perform analysis of the data samples.

At the LHC during the first years of operations, samples of reasonable size were obtained at 7 TeV (in 2010 and 2011), accounting for  $6 \text{ fb}^{-1}$  of luminosity for proton-proton collisions and  $170 \mu\text{b}^{-1}$  for lead-lead collisions. Additionally, data at 8 TeV were obtained for proton-proton collisions, accounting to  $23.3 \text{ fb}^{-1}$ , and proton-lead collisions with a luminosity of  $32 \text{ nb}^{-1}$ . The results described in this report have been obtained by using these data samples.

The collisions provided by the LHC occur at four interaction points along the 27-km ring. At those points, several experiments are located. The main four experiments are ALICE, ATLAS, CMS and LHCb and are located as shown in Fig. 1. These four experiments collect the data from the collisions and provide the results of the physics analyses, as described in the following sections.

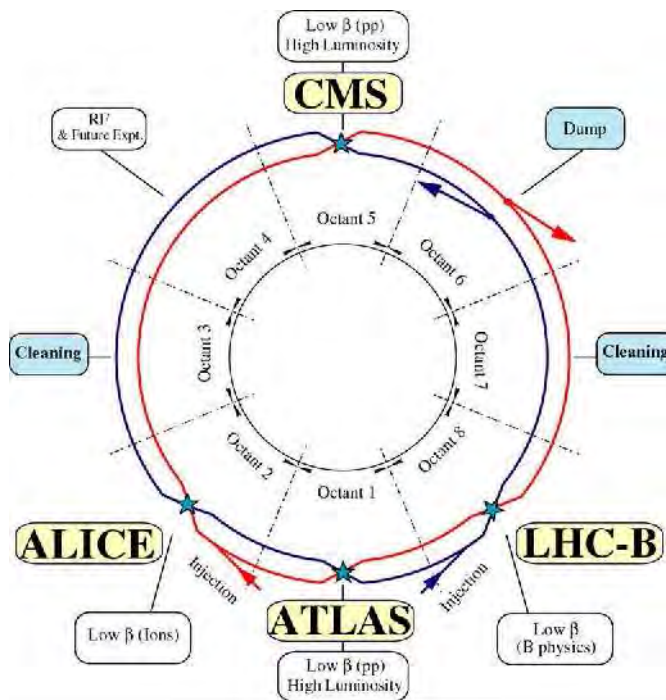
In addition to the main experiments, other three *minor* experiments are intended for more dedicated studies: TOTEM [6], LHCf [7] and MoEDAL [8]. Neither their results nor plans will be covered here since their scientific output is very specific and beyond the aim of this report. However, this should not minimize their importance in order to understand forward production (as it is the case of the first two) or dedicated search for magnetic monopoles (as it is the aim of MoEDAL).

Each of the main experiments deserved some specific description to put into context the physics output they provide.

### 2.1 The ATLAS experiment

ATLAS [9] is the largest experiment at the LHC. It is intended to study all possible physics topic by analysing the full final state of the LHC collisions. It is characterized by its great capabilities in tracking and calorimetry surrounded by huge muon-detection chambers in a toroidal field.

The detector has almost full solid-angle coverage with a forward-backward symmetric distribution. It is also azimuthally symmetric, as expected for the physics in the collisions. The hermetic design allows



**Fig. 1:** Schematic layout of the LHC and the main experiments, identified at their location in the accelerator ring.

to infer the presence of undetected particles via the transverse momentum embalance, the so-called “missing  $E_T$ ” ( $E_T^{\text{miss}}$ ), which can be computed as:

$$E_T^{\text{miss}} = \sqrt{\left[\sum p_x\right]^2 + \left[\sum p_y\right]^2} ;$$

where the sum runs over the observed particles (regardless on the way they are detected and reconstructed).

This quantity is expected to be small due to the conservation of the momentum and therefore a significantly large value is interpreted as the presence of particle(s) that escape detection, as if the case of neutrinos and other weakly-interacting particles which do not interact with matter by mean of the nuclear or electromagnetic forces.

In order to quantify the coverage of the detector, another interesting variable is the pseudorapidity, an alternative to the polar angle  $\theta$  defined as:

$$\eta = -\ln[\tan(\theta/2)] = \frac{1}{2} \ln\left[\frac{|p| + p_z}{|p| - p_z}\right]$$

which is well suited for cylindrical description of events, as it is the case of collisions involving hadrons in the initial state.

The structure of ATLAS allows to reconstruct jets up to  $|\eta| \sim 4.5$ , muons up to  $|\eta| \sim 3$  and electrons and photons up to  $|\eta| \sim 2.47$ , providing a very large coverage for the main pieces to study the final states in the LHC collisions.

## 2.2 The CMS experiment

CMS [10] is the other multipurpose detector of the LHC. Similar to ATLAS in aim and capabilities, it present a more compact structure for a similar performance due to its stronger magnetic field. It is

also hermetic and provides an impressive energy resolution for electrons and photons, for a coverage of  $|\eta| < 2.5$ . Muons are detected up to  $|\eta| \sim 2.4$  with a more traditional approach that takes advantage of the redundancy with the inner tracking. Finally jets are reconstructed up to  $|\eta| \sim 4.5$ .

When comparing both detectors, the strong point of CMS is the great resolution in the inner tracking, which becomes the core of the detector, specifically when used as redundancy for reconstruction of muons and other particles. On the other hand, ATLAS has better global calorimetry and more precise and sophisticated muon detection.

However, these differences are in practice more technical than real, since the treatment of the data in the reconstruction of objects allows both collaborations to obtain very comparable results. The idea is to compensate the limitations of the detectors with the information coming from the stronger parts or redundant informations from other components.

One good example of this is provided by the concept of *particle flow* that has been extensively used in the last years, specially in the CMS analyses. The idea is that instead of reconstructing the event quantities from the detector information (calorimeter cells, tracks), an intermediate step is taken and that detector information is combined to identify “objects” that are associated to particles. From the detector information, the kinematic reconstruction of each “object” is performed in an optimal way, since each class of object (lepton, photon, neutral or charge hadron and so on) is treated differently. It is then from these “objects” that the event quantities are then reconstructed.

These idea represent a big gain since each object is treated as close as possible to its expected behaviour with the detector components. Additionally, the combination of the detector parts allows to get the most of the detector information as a whole, leading to the final goal of having a global event description. The case of CMS is extremely clear since the particle-flow approach allows to use as much tracking information as possible, reducing the impact of the lower quality hadronic reconstruction in the calorimeters.

By the use of this kind of ideas and even more sophisticated techniques, the LHC experiments have been able to extract the most of the data samples, going beyond the most optimistic expectations, as we will describe in future sections.

### 2.3 The LHCb experiment

The LHCb detector [11] has been designed to perform studies on flavour physics, specifically of hadrons containing bottom quarks. Since their production is specially large in the forward region, the detector design is mostly oriented to maximize rate and provide very accurate reconstruction instead of maximizing the coverage. It therefore detects particles in the forward region and it reaches an impressive track and vertex reconstruction due to dedicated sophisticated components.

The main limitation of the measurements in the forward region is the high sensitivity to processes in which multiplicities are large. For this reason, the LHCb did not collect lead-lead data and required *luminosity leveling* to keep the number of collisions in the same event at reasonable levels. This leveling is the reason why the integrated luminosity of the data samples is smaller for this experiment.

On the other hand, its great coverage in the forward region allows this detector to perform measurements beyond the coverage of ATLAS and CMS, providing a nice complementarity at the LHC that is not limited to the topics for which the LHCb was intended. As we will see below, the LHCb experiment is providing nice and competitive results in areas where CMS and ATLAS were expected to be dominant.

### 2.4 The ALICE experiment

The ALICE detector [12] has been designed to maximize the physics output from heavy-ion collisions. The aim of the experiment is not the detection of exotic or striking signatures but to maximize the

particle identification in order to retrieve as much information as possible about the properties of the medium created in the collision and how it affects the behaviour of the produced particles. Therefore, the detector components mostly focus on measurements that allows to study the dependence of statistical properties of the final states with respect of variables that correlates with the production of new matter states, i.e. the production of high energy density, high temperature and high pressure states.

Due to this, the strong point of ALICE with respect to the other LHC experiments is the impressive particle identification, in order to identify relevant particles immersed in high multiplicity events. The limitations that this impose is the reduced coverage for each type of particle and the lack of symmetry in the detector: more types of different subdetectors covering different solid angle regions. This makes that the muon coverage is limited to the forward region ( $2.5 < \eta < 4$ ) while electrons and photons are detected centrally ( $\eta < 0.9$ ).

The specific design of the ALICE detector makes the results from ATLAS and CMS also very attractive for heavy-ion physics, due to its complementarity to ALICE, although they are not in competition when the particle identification is a key part of the study, as we will discuss later in this report.

## 2.5 Data acquisition and event reconstruction at the LHC experiments

The data-acquisition (DAQ) systems of the experiments have been designed to collect the information of the collisions happening at the LHC. They are very sophisticated in order to efficiently collect the information from all the detector components and store it to tape for future analysis.

On the other hand, the DAQ need to deal with the problem that having collisions every 50 ns (or 25 ns in the future) it is impossible to store all or even part of the information for every single event. For that it is needed to have an automated decision system which selects the events as soon as they are produced in order to reduce the amount of data that is physically stored to a manageable level. This system, called *trigger*, has therefore the goal of reducing the rate from tens of MHz to hundreds of Hz, providing data of 100 MB/s, which is a storable quantity.

Although the concept is simple, it should be noticed that events that are not accepted by the trigger are lost forever, implying a big responsibility. Additionally, the trigger conditions at the LHC are very challenging and represent a new frontier in data acquisition due to high rates and event sizes. However, there is the need for those required rates and event sizes since the aim of the experiments is to study rare processes with high precision, even at the cost of suffering at the DAQ level.

In addition to the DAQ challenges, other difficulty arises from the high rate: since the collision cross section is so large, it is very likely that several proton-proton pairs collide in the same event (i.e. crossing). Most of the collisions are soft uninteresting collisions that would appear at the same time as interesting ones. This situation is usually referred as *pile-up* collisions and it complicates the reconstruction of interesting events since it becomes harder to distinct them from usual background, something that is specially dramatic at the trigger level. The reason underneath being that reconstructed quantities, specially the global ones like the  $E_T^{\text{miss}}$ , are modified and led to misleading values.

This problem with the *pile-up* is what motivated the luminosity leveling at the LHCb interaction point: to avoid the deterioration of the performance due to the overlap of collisions. Since statistics is not really the issue due to the large cross section, it is more practical to reduce the collision rate to collect higher purity events than just reject good events due to trigger limitations. It should be noticed that a similar idea may be required for the other experiments in the future when running at the highest rates.

After the data has been collected and stored in tape, it is analyzed to investigate the characterization of the physics producing it. The analysis consists on the identification and quantification of the objects contained in the event.

We have already described how to reconstruct the  $E_T^{\text{miss}}$  quantity that allows to associate undetected particles to the event. Additionally we also described how the reconstruction of the final state may be simplified with the use of the concept of *particle flow*.

As a specific case of the later, the presence of leptons in the final state is a fundamental tool in a hadron collider to recognize important physic events. Electrons are identified using the properties of its interaction with the calorimeters. Muons are identified using the chambers specifically designed for its detection, using the property that they are charged and highly penetrating.

Photons are also identified using the deposits in the calorimeters, where they look similar to electrons, but are distinguishable from them due to the absence of electric charge, and therefore the lack of hits in the tracking system.

The  $\tau$  leptons are the hardest objects to identify in a detector, but their use is strongly motivated by their common presence in final states for BSM physics, or for Higgs searches, as we will see later. Their leptonic decays are hard to distinguish from electrons and muons, but their hadronic decays, the dominant ones, are separated from other hadron production due to their low multiplicity and the kinematical properties. The main issue is that is commonly hard to separate them from the large background of hadron production, and specially at the trigger, where the usable resources are more limited. On the other hand the experiments at LHC has used experience at previous colliders to really exploit all the possibilities of analysis with  $\tau$  leptons, as it is described below.

Finally, apart from leptons and photons, it is very common the production of hadrons. They are originated from quarks and gluons that are not observed because the strong force confines them within colourless hadrons. The mechanisms transforming those coloured particles into hadrons cannot be understood in the perturbation approach used to perform estimations from the theory, but fortunately they can be treated in such a way that their effects do not affect too much the predictions. The simpler technique to reduce this effect is by using *jets* of hadrons to reconstruct and characterize the final states.

The idea is that the processes that are not perturbately calculable occur at energy scales that are much lower than the usual hard processes taking part in the LHC collisions. Therefore they do not modify substantially the global topology of the event and hadrons appear as collimated bunches of particles that are kinematically close to that of the hard partons produced in the event.

This qualitative description, only valid for studies of hard parton production, should be quantified with the use of a specific and well-suited algorithm that reconstruct the jets. The results are usually dependent on the algorithm, but when the same algorithm is used for comparing measurements and theory, the conclusions are independent of the algorithm, if the application is sounded.

Data analyses at the LHC experiments are performed with all these objects: leptons, photons,  $E_T^{\text{miss}}$ , hadrons and jets, with very satisfactory results, mostly due to the high quality of the data acquisition and reconstruction.

### 3 Measurements to rediscover the SM

As mentioned above, the aim of the LHC is to produce unknown particles and increase sensitivity to new possible interactions by colliding protons at high energies. However, on top of the possible interesting processes there are other SM-related processes that tend to hide the most interesting ones. For a hadron collider, QCD jet production has a so large cross section that is the basic process happening in the collisions.

In fact, this makes the LHC a QCD machine aiming for discovery. Independently of what is actually done, everything depends on QCD-related effects: parton radiation, parton distribution functions (PDFs) of the initial-state protons, hadronization processes for the final-state partons and so. Unfortunately most of these cannot be calculated due to our limited knowledge on how to deal with the QCD theory and therefore, in order to understand them requires the realization of measurements which allow to refine the existing phenomenological models used to obtain predictions on what to expect in the proton-proton collisions at the LHC.

For this reason it is impossible to simply ignore the “less interesting” events which are considered as background of the events containing effects and particles beyond the SM. In fact, at the LHC, as in



any other hadron collider, the understanding of QCD is not just something needed nor a priority: it is the only possibility.

As a good example, it is needed to realize that the first measurements performed at the LHC are the total cross section and the differential cross sections for producing charged particles. They are not calculable in the perturbative approach of QCD, but they are required to perform realistic predictions (via the *tunings* of the model generators). They were performed at the beginning of the collisions by all the experiments (see e.g. [13, 14]) and from the beginning have become important tools to understand the collisions at the several energies the LHC has been operating.

In addition, even in these preliminary studies the LHC experiments proved that the LHC is crossing the lines to a new regime: an interesting effect observed looking at the correlations between charged particles: CMS observed [15] that in addition to the usual *large*  $\Delta\phi$  correlations (i.e. opposite hemispheres), there are additional *near-side* (i.e. small  $\Delta\phi$  and large  $\Delta\eta$ ) correlations in events with very high multiplicities, specifically with more than 100 produced charged particles.

Figure 2 shows the mentioned observation of the so-called “ridge”. Similar effects were observed previously in heavy-ion collisions, although it is not completely clear the source of them is the same. Currently there is not a clear explanation of the source, but the LHC data has confirmed its presence in lead-lead and proton-lead collisions, see e.g. [16].

### 3.1 Studies of jet production at the LHC

Apart from these soft-QCD measurements that are a fundamental piece to adjust the phenomenological models, measurements related to hard QCD are also performed at the LHC experiments in order to validate the QCD expectations on the perturbative regime, and to learn about the interactions between partons at the shortest distances and also about the partonic content of the proton.

Measurements are done for inclusive jet production, as those by ATLAS in [17], and compared to the NLO predictions, which are able to reproduce the data after soft-physics corrections (that are not large). Some kinematic regions are sometimes off, but they are correlated to problematic areas, in which proton PDFs are not well known or the effects from higher orders or soft physics are large. Similar conclusions are drawn from studies of multijet production, in which the sensitivity to QCD is enhanced using ratios, as the three-to-two jet ratio by CMS [18], in which many uncertainties cancel and the sensitivity to QCD shows up via the emission of hard partons. In fact the direct sensitivity to the strong coupling constant,  $\alpha_S(Q)$ , allows a measurement of this value for the first time beyond 400 GeV, confirming the expectation from the running of that coupling.

With a different aim, instead of measuring quantities that are more accurately known, there is interest in measuring in regions where uncertainties may be larger, but sensitive to unknown quantities, as it is the case of the PDFs. Measurements at the LHC experiments [19, 20] are sensitive to PDFs in regions where they are not well constrained and able to distinguish between prediction of different sets. Specially useful for the high- $x$  gluon and sea quark PDF which is loosely constrained by the HERA data. It is worth to remark that even if the LHC aims for discovering of BSM physics, it is a very useful machine to increase the knowledge about the internal structure of the proton, via the measurements sensitive to the PDFs. In incoming sections this will be mentioned a few times.

When studying the production of jets, an important topic by its own is the measurement of production of heavy-flavour (charm and bottom) jets. Since they are not present in the proton in a sizable way, its study provides important information about QCD, specially for specific flavour production, something which is not possible for the light quarks and gluons. The fact that it is possible to perform separated studies for charm and bottom jets is due to the possibility of tagging the jets as originating/containing a heavy-flavour quark.

This has been a recent possibility due to the improvement in tracking, specifically at the closest distances to the collision. After surpassing the challenges involved in the LEP and Tevatron experiments,

the detectors have reached to possibility to reconstruct vertices so precisely, that resolving secondary vertices coming from “long lived” hadrons containing a bottom and a charm quark has become a standard tool in accelerator physics.

The fact behind this *heavy-flavour tagging* is the presence of hadrons that live long enough so their decay products appear in the detector as displaced tracks and vertices within jets that are incompatible with originating at the so-called primary vertex, in which the interaction took place. These displaced tracks and vertices are resolved and conveniently used to tag jets containing these heavy-flavour hadrons and therefore likely to originate from a charm or bottom quark. The information provided by them is used either on a simple and straightforward way (that is safer and more traditional) or on multivariate techniques that allow to increase performance of the tagger. The later has become more popular as expertise with this kind of tool is well established.

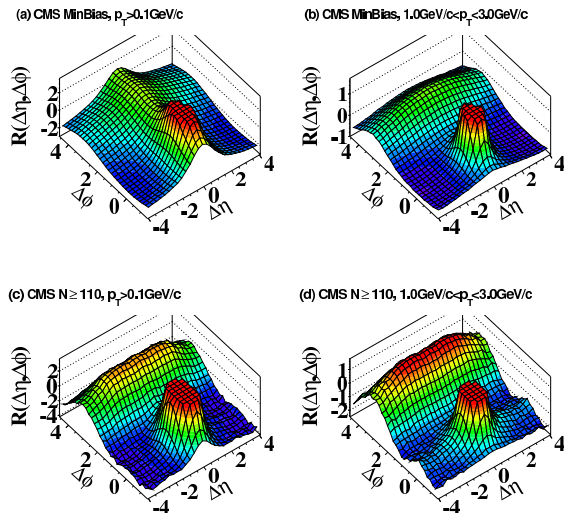
Making use of the tagging tools it is possible to study the production of jets originating from a bottom quark, or b-jets. Measurements by the two collaborations has been made [21, 22] and compared to QCD precisions for heavy-flavour production computed with the MC@NLO [23] program. As shown in Fig. 3, a good agreement is observed overall although there are some small discrepancies in specific kinematic regions, similarly at what was observed in inclusive jet production. It should be noticed that the level of agreement is good due to the improvements in the theoretical calculations during the last decade. Predictions are difficult for the kind of process under study, so the level of discrepancy observed is considered a complete success of the QCD calculations. Of course further work is still needed, emphasizing the importance of the precise measurements at the LHC.

In a similar topic, one important measurement at the LHC experiments will be to try to disentangle the production of jets containing two heavy-flavour quarks. In the past the quality of the heavy-flavour tagging only allowed the separation of jets with at least a heavy-flavour quark. However, at the LHC, the improved detection techniques and the experience with tagging tools will also allow to investigate the production of multi-b jets, which are of importance in topologies with merged jets or to reject the presence of gluon jets containing a gluon-splitting process into heavy-flavour quarks.

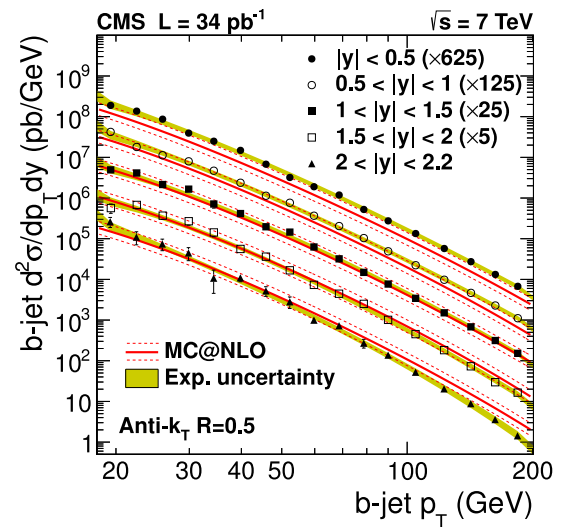
Exploiting the subtle differences in the displacement of tracks, studies are performed on this issue [24], and good rejection power of gluon jets has been observed while keeping a big fraction of the single b jets. More dedicated studies will be needed to improve the related tools for rejecting this background, but current results has confirmed its feasibility and also that the heavy-flavour taggers at the LHC experiments are taking advantage of the improved detector capabilities.

Regarding the LHC in a new kinematic regime, it should be remarked the development during the last years of tools to investigate the production of boosted objects. Since available energies at the LHC are much larger than the masses of the SM particles, it is likely to observe their production with very large transverse momenta, giving rise to the merging of objects. This is specially worrisome in the case of jets since they are hard to separate after their constituents have been merged together. For that reason, several dedicated studies and the development of new techniques has been done at the LHC experiments [25, 26] in order to deal with the topology of boosted jets. The idea is to exploit the properties of the internal structure to recover information of the original partons whose jets have been merged, and separate them from single parton jets that are boosted in the transverse direction, i.e. produced with large transverse momentum.

Many techniques have been developed and tested in the identification of merged jet and check how the simulation reproduce the characteristics of the jets allowing the distinction of the jets containing one or more “hard” partons. Currently its performance has been proven to identify merged jets coming from boosted W bosons and top quarks, and used for searches. However its principal motivation is still the need of this kind of tools for the future running at higher energy.



**Fig. 2:** Relative distribution of the charged particles in proton-proton at 7 TeV as measured by CMS in several selections in the  $\Delta\eta - \Delta\phi$  plane. Apart from the expected back-to-back correlation, a near-side correlation is observed even at large  $\Delta\eta$  for high-multiplicity events (plots below).



**Fig. 3:** Differential cross section of bottom jet production at the LHC with a center of mass energy of 7 TeV. Measurements in different rapidity regions (dots) are plotted as a function of the transverse momentum of the jet and compared to the MC@NLO predictions (lines).

### 3.2 Studies of soft QCD physics at the LHC

Apart from particle and jet production via QCD processes, the experiments are able to perform studies related to QCD via more complicated mechanisms. Among this, one that has become really important is the possibility of observing more than one partonic collision from the same protons. Since a proton is a bunch of partons it is not uncommon to have several partons colliding at the same time. And the LHC allows to have very hard collisions since the energy of the protons is very large.

These multiparton interactions are a complicated topic since it is not clear up to which level each collision can be considered independent of the others. In addition, the probability associated to the additional collisions to happen is not calculable and require models whose parameters require some tuning in order to improve the modeling of the underlying event. The validity test of the models is usually done in samples that are reasonably understood and trying to extract the maximum possible information to get the proper parameterization. With this aim, ATLAS has measured the contribution from double-parton interaction for W+dijet events [27] to be  $0.16 \pm 0.01(\text{stat}) \pm 0.03(\text{syst})$ , in good agreement with the expectations that were tuned to previous data.

Related to QCD in strange regions, the LHC allows studies for diffractive and forward production of particles and jets at higher scales than previous hadron colliders. These are relevant in order to understand hadron interaction at softer scales, and also to adjust the models describing this kind of process.

Even the LHCb experiments has produced results for forward hadron production, which are very competitive due to the optimization of the detector for particle ID and its very forward coverage. Results of these studies [28] have been compared to the predictions obtained by traditional event generator and also those used in the simulation of cosmic-ray events, which are very sensitive to this kind of processes.

Another example of new kind of QCD measurements is the study of exclusive diboson (WW) production via the collision of photons performed by CMS [29]. This makes the LHC a photon collider at high energies, which allows dedicated studies of the electroweak interaction. The result with the

dataset collected at 7 TeV allows to measure the cross section with still a low significance, implying the need of more data. However, using the sample with highest transverse momentum, it was possible to set limits on the production via anomalous quartic couplings, showing the potential of this kind of studies.

### 3.3 Electroweak boson and diboson production at the LHC

Although the measurements described above allow to test the predictions by QCD and even of the electroweak sector in some cases, the most sensitive studies to validate the SM predictions are coming from the events containing photons or weak bosons. The idea is that these events are usually simple to recognize and the perturbative calculations of the processes and the backgrounds are usually very accurate.

The most common process of this kind is the production of photons, whose interest have been demonstrated in the past hadron colliders, in which this was considered a “QCD study” since it provided direct information on the quarks. Hard photons radiated from quarks are good probes of the interaction since they are not affected by soft processes and they are able to distinguish among different kind of quarks. In addition the large cross section of the  $\gamma$ +jet allows its use as a fundamental calibration tool.

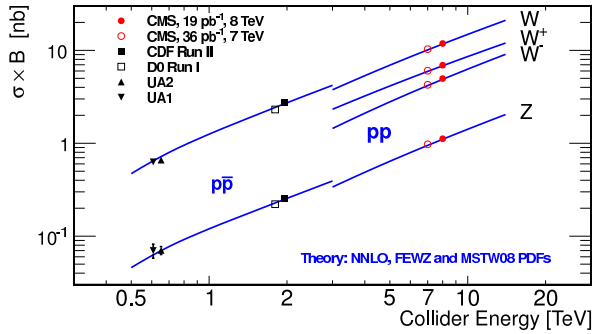
Additionally, studies of diphoton production yield to very stringent test of the SM predictions, specially for a final state that is an important background in many interesting searches of new particles, decaying in photon final states. The study by ATLAS [30] performed measurements of the photon pair production as a function of several variables and compared them to several event generators, at different orders in QCD and types of partonic showers in order to evaluate the level of performance of the available production tools.

However, when talking on boson production, the studies related to the weak bosons become a fundamental test of the SM predictions that were performed at the LHC in order to also check the performance of the detectors and tools for analyses. Even after the first analyses, the studies of events with W and Z bosons are fundamental tools for calibration and understanding of the object identification and reconstruction. Measurements at several energies, as the one at 8 TeV by CMS [31], have been performed and show very good agreement with the expectations by the SM and also confirming the excellent predictions of the SM at several energies for measurements performed for W and Z production during the last three decades, as shown in Fig. 4.

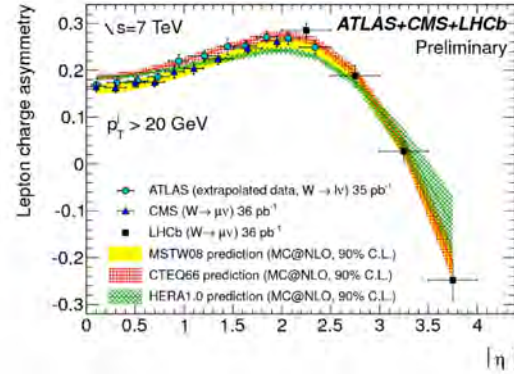
Although the basic goal for studying the production of weak bosons is to confirm the performance of the detectors and of the basic SM prediction, dedicated measurements related to them are also a fundamental part of the LHC program. This is the case for measurements sensitive to the internal structure of the proton and also of the SM details that could not be tested before at the level of precision reachable at the LHC. This affects both kind of processes: final states that were never available in a proton-proton collider before, like the ratio of  $W^+$  to  $W^-$  measured by ATLAS [32], or whose yield was too small, like the measurement of  $Z \rightarrow 4l$  (as in [33]) which is a calibration piece for the Higgs searches.

This explains the large effort at the LHC to measure the properties of the production of weak bosons. Some of the properties are measurements for confirmation and validation purposes, but some are really motivated by the new possibilities opened at the LHC experiments. This is seen even in experiments that are not intended for boson studies, like the results at LHCb, in which the very forward detection makes measurements of Z and W production very competitive even with lower acceptances [34], since they are measured in kinematic regions that are not available for the main detectors. Even events compatible with forward Z bosons decaying into  $\tau$  leptons have been observed at the LHCb [35], indicating an important benchmark for the performance of the experiment to obtain results beyond flavour physics.

In the case of W production, Fig. 5 shows the lepton charge asymmetry as a function of  $\eta$  also confirms the complementarity of the several experiments at the LHC, in this case how the LHCb is able to extend the region reachable by the ATLAS and CMS, even with a reduced yield. All these measurements of forward production will have a big impact in the fits to extract the parton content of the proton, since



**Fig. 4:** Cross sections of weak boson production in hadron colliders at several center-of-mass energies. SM predictions for proton-antiproton and proton-proton collisions are compared to the measurements shown as different types of dots.



**Fig. 5:** Lepton charge asymmetry of W production as a function of  $\eta$  at the LHC with the 7-TeV data from the ATLAS, CMS and LHCb experiment. The inclusion of the later experiment allows to extend the measurement to very forward regions never reached before.

most of the current uncertainty is reduced by forward production of particles, more sensitive to the less constrained partonic content, as gluons and sea quarks at high- $x$ .

But not only the proton structure benefits from the large yields at the LHC for producing weak bosons since the presence of a massive object allows studies of QCD processes in an environment where perturbative calculations are accurate enough to bring very stringent tests of the expectations.

The typical example is the use of bosons as “probes” of the underlying hard process involving the partons, whose rules are naturally dictated by strong interactions. This is the case of the measurement of jet production in association to a Z or W, as in [36, 37], which are sensitive to the partons interacting and also major backgrounds to most of the new models for BSM physics. The measurements are able to constrain the room for the new physics, and, in other kinematic regions, to check the validity of the tools used to estimate these final states. It should be noted that not only the yields are interesting, but also the kinematic distributions of the final state objects, specially those sensitive to unexpected underlying physics, as in [38, 39], in which specific distributions of bosons and jets are studied in order to perform accurate tests of the SM predictions, taking advantage of the large yields.

Similarly, another topic that directly benefits from the high cross section and luminosity at the LHC is the production of heavy flavour quarks in association with a weak boson. Being very sensitive to the SM structure, some of the processes have not been accurately tested due to the limited statistics at previous colliders. In fact, results at the Tevatron have been controversial regarding the way the event generators reproduce the measurements. The larger statistics at the LHC allows the improvement in the precision of the measurement. This is the case for the W+b-jet measurement by ATLAS [40], which clearly shows that description by event generators could be improved, which is not a trivial case, since it is a background for many studies for BSM physics. Understanding this discrepancy should be a clear priority of the physics at LHC, from the theoretical and experimental point of view.

Another final state that has benefited a lot by the new frontier set at the LHC is the production of charm in association with a W boson. Its interest is given by the fact that since W is able to change the flavour of a quark, the production of single charm is dominated by interactions involving down and strange quarks in the proton. Therefore directly sensitive to the strange content of the proton. In addition, the charge of the produced W is completely correlated to the charge of the charm and down/strange quark. As mentioned above, the W is used as a direct probe of the structure of the underlying parton collision. In this case the result of the measurement by CMS [41] is presented as the fraction of charm jets in W+jet events and also of the ratio of  $W^+$  to  $W^-$  in events with a charm produced in association with the W.

Both quantities are sensitive to the PDF of the strange quark and antiquark. The measurements are in good agreement with the expectations and they will allow to improve the accuracy of the proton PDFs.

In the case of the Z boson, the low cross sections prevented detailed studies of the production associated with heavy flavour quarks to be performed at the Tevatron. Again the LHC has brought the possibility to study this in detail. The analyses studying the production of Z+b-jet, as in [42], show that the event generators, in this case MADGRAPH [43], are able to describe the distributions. However, with the explicit requirement of two b-jets the agreement get clearly worse [44], implying that some theoretical work may be required: although the processes (and calculation diagrams) are the same, the relative weight is different due to the kinematic requirements on the second jet.

Finally, the last topic entering the scene when talking about weak bosons and jets is the study for electroweakly produced bosons, the so-called *Vector-boson fusion* (VBF) production. In this case the boson is produced in association of two jets that tends to be forward, due to the kinematics. Those forward jets are used to “VBF-tag” the event and separate them from the main processes, weak radiation from partons or parton annihilation. Measurement by CMS [45] allowed to measure a cross section in agreement with NLO calculations. In addition, this kind of analysis also contributes to understand the production of jets in the forward region, which is less understood due to the challenges in experimental studies and also in theoretical calculations.

It should be remarked that the interest of all the results involving jet production in association with weak bosons will be kept in the future, as the measurements get more precise, implying larger challenges for the modeling of very important processes at the LHC, either for their own interest or just as background estimations for searches of all kind.

### 3.4 Diboson production at the LHC

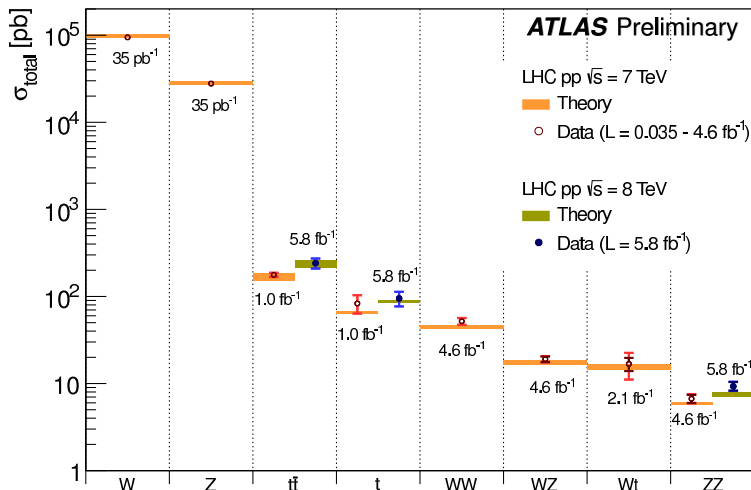
As it is well known, the production of more than one boson is one of the most sensitive test of the non-abelian structure of the electroweak sector of the SM, so it is very sensitive to deviation produced by new couplings involving the SM bosons.

The main limitation is that precisely the presence of several weak couplings makes the cross section small, and the observation of these final states has been very difficult. However, the LHC has open a new era for this kind of studies since large samples are available to perform detailed studies, allowing precise studies of diboson production for the first time. In fact, the LHC will allow in the future the observation of multiboson production, which has never been observed. In addition, the large samples available has allowed that diboson production has become a standard reference for calibration in advanced analyses.

The basic processes testing the SM structure and with large cross section is the production of a weak boson and a photon ( $W\gamma$  and  $Z\gamma$ ) which are directly sensitive to the unification of the electromagnetic and weak interactions. The results of the analysis, like [46], shown that data are in good agreement with expectations, even at higher transverse momenta, which may be sensitive to new physics affecting the unification of interactions.

In the case of two massive boson, the process with the highest cross section is the production of two W bosons, in which the samples are large enough to allow detailed comparisons with the predictions by the event generators, even via differential distributions [47]. The conclusion of the studies is that the SM predictions reproduce very well the shapes of the observed distributions in data, but they underestimate the total cross section.

This discrepancy has been observed by the two collaborations and at the two energies of the LHC. Investigation of the origin of it is under study. Similarly, studies of the production of two Z bosons shows a slight excess in the data with respect to the expectations [47, 48]. In this case, the yields are small and the excess is not as significant, but the clean final state, requiring four isolated leptons, leads to very straightforward conclusions. This channel, which leads to a pure sample of ZZ events and with fully



**Fig. 6:** Summary of the measurements by ATLAS for massive particles (weak bosons and top quarks) in single and double mode at the LHC with a center-of-mass energy of 7 TeV.

reconstructed kinematics, provides the best test bed for diboson studies, specially with the amount of events expected at the LHC.

In addition to the pure leptonic channels, that are much cleaner in a hadron collider, the semileptonic channels are also exploited at the LHC, since it is the most precise way to study the hadronic decays of Z and W bosons, not available in the inclusive production due to the large dijet backgrounds. The performed measurements in the W+dijet sample [39, 49] yield the observation of the diboson signal. Separation of the Z and W in the hadronic channel is not possible due to resolution, and therefore this final state is able to measure the mixture of WW and WZ events. The result is in agreement with the observation, and the analysis has also tested the W+dijet background, whose interest was mentioned above. Finally it should be remarked that WZ has been also measured in the fully leptonic channel [50] which provides the topology of three charged leptons and  $E_T^{\text{miss}}$  which has a large relevance in searches for new physics, in particular supersymmetry, and therefore the understanding of the kinematics in this diboson process is a fundamental part of the program.

In conclusion, it should be remarked that even if the LHC is intended to discover the physics beyond the SM, measurements of the know processes has produced many interesting results, some to confirm the observations at previous colliders, but also new results that were not previously accessible. In this sense, and as summarized in Fig. 6, the impressive agreement of the measurements provides a solid base on which the experiments are building the tools and confidence for the observation of unexpected results, when higher precision or new final states are reachable in the data.

#### 4 Measurements on bottom and charm hadrons

The spectroscopy of hadrons has been a fundamental source of information in particle physics, since it has allowed to detect effects beyond the reachable energy scale and since it provides the only direct way to understand quarks and QCD at low energies.

The case of heavy flavour hadrons, which include at least a bottom or charm quark, is of a broader interest due to the higher masses involved that allows to perform more accurate theoretical calculations related to the properties of the hadrons. With the measurements in hadron spectroscopy, it is possible to perform several classes of studies, as the properties of bound states, production of new states, measure branching ratios and interference effects. All of them provide information about possible BSM physics or improve the knowledge about partons in confinement states.

It should be remarked that in order to perform studies with hadrons, it is needed to reconstruct them. This sets a very different approach to the ones described above in which the hadrons are just merged together in jets that are related to the original partons. The goal in the physics with hadrons is to explicitly identify the interesting objects. This is achieved in several steps: The first consists in the identification of the detected particles, as pions, kaons and more commonly muons and electrons. Some of these objects are (pseudo)stable and are identified as tracks or similar. Sometimes the nature of the particle is also inferred by using specifically designed detectors, but in other cases the nature is just assumed as part of the reconstruction process.

After the detected particles are identified, they are combined to reconstruct “mother” particles that may have decayed into them. The usual method is to reconstruct the invariant mass of several identified objects and find events in which they are coming from another particle (over a possible continuous background) as a resonant excess. Those events associated to a decaying particle may be used to extract information about the particle, apart from the direct identification of the particle itself in the mass distribution. Furthermore, the particles identified this way via its decay products may be further used to reconstruct other parental particles in a recursive reconstruction that allows the full identification of the decay chain of the original particle.

With these tools and the goal of measuring the hadron properties in mind, the LHC experiments have been able to identify hadrons, some of them completely unknown. One example is the observation by ATLAS of the new excited state,  $\chi_b(3P)$ , belonging to the bottomonium family decaying into  $\Upsilon(1S/2S)$  by the emission of a photon [51]. The mass distribution showing the resonances produced by the new state is shown in Fig. 7 centered at a mass of  $10.530 \pm 0.005(\text{stat}) \pm 0.009(\text{syst})$  GeV. Also the CMS experiments was able to find the  $\Xi_b^* \rightarrow \Xi_b^\mp \pi^\pm$  state, which has been the first baryon and fermion found at the LHC, and with a mass of  $5945.0 \pm 0.7(\text{stat}) \pm 0.3(\text{syst}) \pm 2.7(\text{PDG})$  MeV [52].

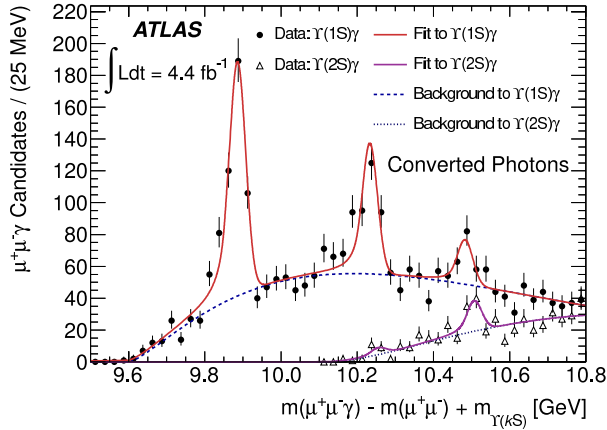
However, and as expected, it is the main experiment focusing in heavy-flavour physics, LHCb with its larger samples with higher purity who is able to measure the properties of bottom hadrons with higher precision. Specially about the recently discovered baryons, for which this experiment has already relatively large samples with high purity selection. The measurements for  $\Lambda_b$ ,  $\Omega_b^-$  and  $\Xi_b^-$  documented in [53] required very detailed understanding of the detector momentum scales, in order to get the most precise mass measurements in the World.

Additionally the LHCb is also leading the effort in searching for rare decays of known hadrons. These decays are of interest for its possible sensitivity to new interactions involving quarks because they include loop diagrams or interesting vertices that could be affected by unknown effects. Among the rare decays, one of the most attractive ones is  $B_s/B^0 \rightarrow \mu\mu$  since it is associated to a well-controlled and easily identifiable final state. Additionally, the branching ratio is very small but expected to be enhanced in several of the possible BSM extensions. This explains the intensive search for this signal in the last decade at the Tevatron, where exclusion limit approached the SM expectation. However, the large sample collected by the LHCb experiment allowed to get evidence of the decay, with a significance of  $3.5\sigma$ , for  $B_s$  that is in good agreement with the SM value [54]. The decay for  $B^0$ , searched in the same analysis, is also in agreement with the SM, but significance of the excess is smaller. The absence of discrepancy has set strong limits on possible new physics affecting the decay, confirming the negative results from direct searches at the other LHC experiments, as described in sections 8 and 9.

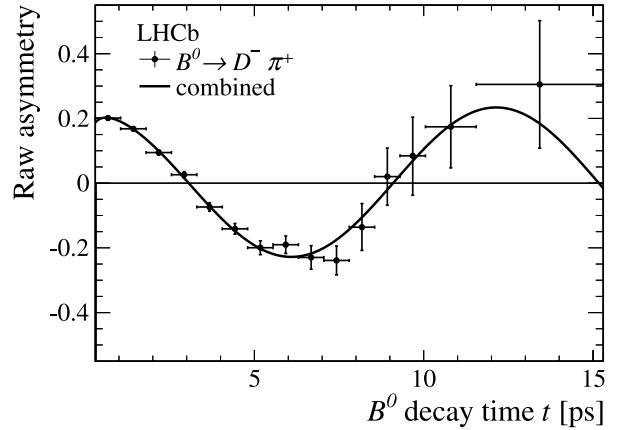
Another interesting decay under study is  $B^0 \rightarrow K^* \mu\mu$ , whose branching fraction in the SM is not that small but whose kinematics is sensitive to the presence of new physics. One is the forward-backward asymmetry as a function of the invariant mass of the muons, measured by LHCb [55] and observed to be in agreement with the SM calculations.

All these measurements confirm the good performance of the detectors for heavy hadron physics, although the measurements are not bringing information about the possible BSM physics, but setting stringent constraints on the way the new physics may modify the interaction between quarks.





**Fig. 7:** Invariant mass distribution of  $\mu^+\mu^-\gamma$  to observe the resonances decaying into  $\Upsilon(1S/2S)$  and a photon. A clear state at 10.5 GeV is observed in both decays, compatible with being the  $\chi_b(3P)$  state of the bottomonium family.



**Fig. 8:** Raw mixing asymmetry for  $B^0 \rightarrow D^-\pi^+$  as a function of the decay time. The solid black line is the projection of the mixing asymmetry of the combined probability function for the sample.

#### 4.1 Mixing and oscillations

Within the properties of hadrons, one that has become of large relevance is that of the mixing of neutral mesons, in which the flavour eigenstates differ from the mass eigenstates, leading to a change in its nature according to the quantum mechanics rules. These oscillations are well established for the  $K^0$ ,  $B^0$  and  $B_s^0$  and are starting to become accessible for the  $D^0$ .

In the case of the  $B^0$ , the LHCb samples are reaching unprecedented precision and even providing new channels of observation. Figure 8 shows the result of the oscillations for the very pure sample of  $B^0 \rightarrow D^-\pi^+$  as a function of the decay time [56]. As it can be observed, the measurements are well reproduced by the expectation obtained taking into account the composition of the sample used to compute the raw asymmetry.

In the case of the  $D^0$ , the oscillations are now becoming accessible thanks to the large samples, specially at the LHCb. Its study is strongly motivated since charm is the only up-type quark in which mixing and CP violation are accessible. It can also provide surprises since it is a previously unexplored region. The study of the mixing and oscillations for the  $D^0$  is done by exploiting the interference between the mixing and the double-Cabibbo-suppressed decays. The same channel provide a right sign and a wrong sign set of candidates that are used to perform the measurement. The first set is not sensitive to the mixing and therefore provides a perfect reference sample.

In order to reduce uncertainties in the production, the initial  $D^0$  state is tagged by using the decay product of the  $D^* \rightarrow D^0\pi_s$ . Using all these events, it is possible to measure the mixing and the LHCb has provided the first observation from a single measurement, with a significance of  $9.1\sigma$  [57]. The result is in good agreement with previous measurements, but the increased significance is another proof of the reach available at the LHC even for studies of low-mass objects.

#### 4.2 Measurements of the CKM matrix and CP violation

As remarked several times, the main goal of the studies in flavour physics is to investigate the details of the fermion families, specially the relationship among them. In the case of the quarks, the relation between the flavour eigenstates (from the point of view of the weak interaction) and the mass eigenstates. is given by the so-called CKM matrix [58] which is expected to be unitary (when all families are included) and that can be parameterized with three mixing angles and one complex phase. The unitary condition

allows the representation of combinations of elements in rows and columns of the matrix as a triangle whose area is related to the CP violation in the family mixing.

The goal is therefore to identify the processes that are sensitive to combinations of elements in the matrix and extract the associated information about the matrix and the triangle. The measurement of single elements in the matrix is associated to processes that are not observable in hadron physics. However, that is not a complete limitation, as proven by the large set of results in the last decade related to the CKM and CP violation parameters. Still, certain measurements are newly coming from the LHC. As an example, the LHCb experiment has measured the angle  $\gamma$  using the tree processes  $B^\pm \rightarrow D^0 K^\pm$  [59] which has the advantage of being very clean: as we mentioned before, processes with loops are sensitive to new physics, so the values measured at tree level are dominated by SM-only physics. The measured value,  $\gamma = (71.1^{+16.6}_{-15.7})^\circ$  is in agreement with the World average, with comparable uncertainty.

Other interesting result from the LHCb is the study of CP violation in charmless three-body decays of B mesons [60], that are sensitive to transitions between the first and third generation. The observed asymmetry is interesting because it is opposite in  $\pi^\pm \pi^+ \pi^-$  (enhancement for  $B^-$ ) with respect to  $K^+ K^- \pi^\pm$  (enhancement for  $B^+$ ) and it seems to be enhanced locally for some kinematics regions.

In the case of the mixing, one of the most important channel is  $B_s \rightarrow J/\psi \phi$  since it is sensitive to new physics affecting the CP violation. Measurements [61] agree with the SM expectations, and they were also used to obtain the first measurement of the width difference of the mass eigenstates which is not compatible with zero ( $\Delta\Gamma_s = 0.116 \pm 0.018(\text{stat}) \pm 0.006(\text{syst}) \text{ ps}^{-1}$ ).

Finally, the last open topic for CP violation is its study in charm decays, which has been measured by the LHCb collaboration [62] to be significantly different from zero, an unexpected result since most of the SM-based predictions suggest almost no violation. Although calculations are difficult and the usual estimations may underestimate the value, the measured value, confirmed at other experiments, seems a bit large, which may be pointing to some BSM effects.

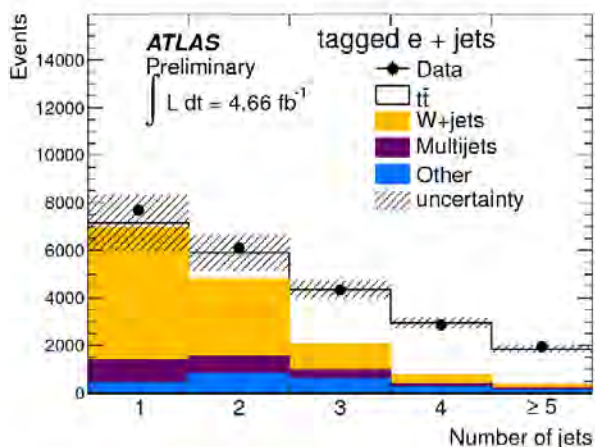
As with most of the discrepancies observed, more data is needed to increase our knowledge, but theoretical development is an additional requirement to quantify the level of disagreement observed and before its origin is further investigated.

## 5 Results on the top quark

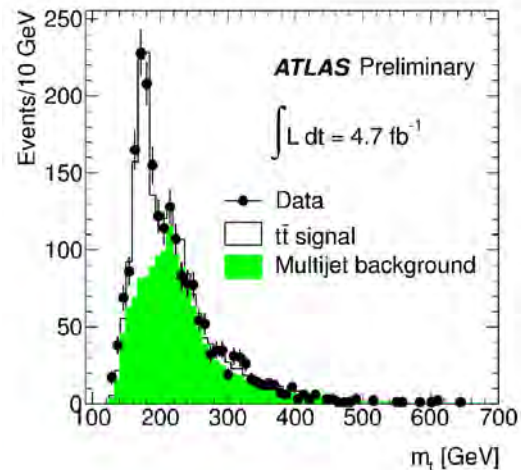
In the hadron physics described in the previous section, one quark is not investigated: the top. Being the most massive of the quarks (and of any observed fundamental particle) it is hard to produce and also it does not hadronize but directly decays into a W and a bottom quark. Additionally, its exceptionally high value of the mass makes him the best candidate to be related to new physics, so its study is mandatory and one of the big goals of the LHC program: the top quark may lead the path to BSM physics, in the same way as neutrinos are leading the path in non-collider results.

At the LHC the dominant process to produce a top quark is QCD pair-production that has a large cross section. In fact the LHC is the first machine that is able to produce top quarks at high rate, allowing detailed studies to be performed. This also applies to other production mechanisms, as that of single-top and  $tW$  production, the latter being available at the LHC for the first time. In fact the production cross sections of processes involving top are so large that it is also a very common background in many types of searches, which is an additional motivation for studying its properties.

The study of the top quark at the LHC follows a similar strategy developed at Tevatron: channels are identified with the number and type of leptons in the final state. Depending on that, events are analyzed to extract all available information in a sample as clean as possible. Additionally all channels are considered, in order to investigate all possible events and the presence of discrepancies with respect to the SM expectations.



**Fig. 9:** Distribution of the number of jets in events with an electron or positron, a b-jet and significant  $E_T^{\text{miss}}$  as measured by ATLAS at 7 TeV. Sample composition is split into the main components.



**Fig. 10:** Invariant mass distribution for three jets forming a top candidate in fully-hadronic of top-pair production events. Measurement by ATLAS at 7 TeV. Expectations for the top-pair signal and the multijet background (histograms) are shown and compared to the data (dots).

### 5.1 Measurements of the top-pair production cross section

The first property to be measured for the top quark is the production cross section in the main mechanism (pair production) and the simpler channel: the semileptonic events in which there is a good identified lepton and at least one jet tagged as coming from a bottom quark. Results were obtained for the sample collected at 7 TeV by ATLAS, giving a cross section of  $165 \pm 2(\text{stat}) \pm 17(\text{syst}) \pm 3(\text{lumi})$  pb [63]. Distribution of the number of jets is presented for events with an electron in Fig.9, showing the clear signal yield for high jet multiplicities.

It should be noted that the semileptonic events apply only to electrons and muons, not to the  $\tau$  lepton that is considered aside. That channel has also been studied since it is very important for the possible new physics related to the third generation and the measurements (like the one in [64]) are found to be in good agreement with the expectations. Additionally the all-hadronic channel has also been investigated [65] in order to confirm the expectations. These two channels used the invariant mass distribution of the top quark candidates, as shown in Fig. 10, in order to separate the large backgrounds. It should be remarked that the lack of precision for these channels is basically driven by the systematic uncertainties affecting the background or the acceptances.

On the other extreme, channels containing two leptons (electrons and muons) provide the cleanest signature. At the Tevatron this channel was not precise because of the lower yield, but the LHC has proven this is no longer an issue with the single most precise measurement of the cross section from the dilepton channel at CMS [66],  $161.9 \pm 2.5(\text{stat})_{-5.0}^{+5.1}(\text{syst}) \pm 3.6(\text{lumi})$  pb, again at 7 TeV.

All these channels provide experimentally independent measurements of the production cross section that have been combined [67] to give a value of  $173.3 \pm 2.3(\text{stat}) \pm 9.8(\text{syst})$  pb. The combination has also proven the good consistency among the different channels and the two experiments. In addition to these results at 7 TeV, the two collaborations are working on getting a similar picture with the data collected at 8 TeV and measure the top-pair production cross section, whose interest is to test the model at higher energies but also to open the possibility of performing ratios of energies (and even double ratios with the addition of the Z-boson production cross section) which will enhance the sensitivity to BSM physics. The first measurements of the cross section at 8 TeV are reported in [68] and [69].

However it should be remarked that the large samples of top events are also allowing new sets of studies that were not available at Tevatron: measuring SM quantities using events containing top quarks. Those provide good tests of the SM, but also a useful frame to perform precise measurements. One example is the extraction of  $\alpha_S$  from the top-pair production cross section [70], which leads to a competitive value because it is determined in an energy regime that has only been accessible to a reduced amount of measurements.

Besides of the total production cross section, the experiments are measuring differential cross sections [71, 72]. These studies provide very stringent test of the SM predictions and of the modeling in simulation. In addition the sensitivity to possible discrepancies is enhanced, since such discrepancies could appear in tails of distributions, as expected from possible new physics, and not affect the bulk of them in any visible way.

The results of the measurements does not present any significant discrepancy and good agreement is observed, which increases the confidence on the predictive power of the theoretical tools. These are going to be fundamental when larger samples are investigated, as those collected in 2012 at 8 TeV, since precision will be much larger and the challenges and sensitivity to new physics increases to previously unknown levels.

## 5.2 Measurement of the properties of the top quark

Until more data is available for detailed studies of the production mechanism, the current data samples allow the measurement of the properties of the top quark to an unprecident precision. The first one is the determination of the mass, since it ia a parameter that determines many other properties, and its high value is already a motivation by itself.

The LHC experiments are exploiting the experience at the Tevatron and are already measuring the mass of the top quark with very advanced techniques: template fits, jet calibration in-situ and similar. In addition the measurements are performed in several samples that are later combined, even to get a combined LHC result, as summarized in Fig. 11 and documented by the collaborations [73]. It should be remarked that the achieved precision will be very hard to improve, but still the mass of the top quark is a relevant quantity of study at the LHC. Specfically larger samples will allow differential measurements of the mass,  $dM_t/dX$ , which provides additional information and constraints.

In addition to the direct measurement of the mass, the LHC experiments are also measuring the mass indirectly from the measured cross section and the comparison to the theoretical expectations. The value extracted from this [74, 75] is not as precise as the direct measurements, but the comparison provides a new handle to find inconsistencies in the theory predictions (and therefore opening the way to possible BSM physics). The results are in good agreement, confirming the impressive performance of the SM predictions for top production and properties.

Additionally to the mass there are other several quantities that have been measured for the top at the LHC by CMS and ATLAS. As an incomplete summary, here are brief references to them:

### – Electric charge

Within the SM there is a fixed expectation for the electric charge of the top quark (+2/3 of that of the positron). However, some models would allow a charge of -4/3 (same units) which is still fully compatible with the observed decays since the inclusive measurements do not relate the charge of the lepton from the W boson and that of the bottom quark, specially due to the difficulties to measure the latter.

However performing studies of the charge asassociated to the bottom quark (and the jet) and the pairing of jet and W boson to identify the ones coming from the same top, it is possible to obtain sensitivity to the charge of the top quark. Even with limited luminosities, analyses by the two collaborations [76, 77] by testing the two models again sensitive distributions are excluding the alternative value beyond any reasonable doubt.

### – Mass difference for top and antitop

CMS has measured the mass difference between the quark and the antiquark version of the top [78], which provides a stringent test of the CPT invariance in Nature and of the possible compositeness of the top quark state. The result is in agreement with the SM expectation in which there is no difference.

### – Polarization and spin correlations

Due to the short lifetime of the top quark, its decay happens before a change of the spin. This allows to perform studies related to the spin that are not available to any other quark.

In pair production the polarization of the top quark is investigated by using the angle between the quark and the lepton. Measurements by CMS in the dilepton channel [79] and by ATLAS in the lepton+jet sample [80] has confirmed that the polarization is in agreement with the SM expectation: top quarks are produced unpolarized.

However, the SM predicts that even if the quarks are not polarized, the spins of que quark and antiquark are correlated. The degree of correlation as measured by ATLAS in helicity basis is  $0.40_{-0.08}^{+0.09}$  [81], in perfect agreement with NLO SM predictions, which sets additional constraints to possible anomalous production, i.e. BSM physics.

### – Helicity of W from top decays

Due to the characteristics of the coupling of the W boson to fermions, we expect that helicity of the W decaying from top quarks to be fully determined. This property is parameterized in different components that are accessible by studying the angular ditributions between the lepton from the W boson and the top quark in the W rest frame.

Measurements performed by the two collaborations [82, 83] are in agreement with the SM expectations and the results are used to set limits on anomalous couplings between the W boson and the top quark, basically testing the V-A structure of the weak coupling of the only quark in which it is directly accessible.

### – Forward-Backward asymmetry in top-pair production

In top-quark pair production a striking assymetry was observed at the Tevatron regarding the foward-backward production of the quarks, which a clear preference of the top quark to be produced in the direction of the proton (and the antiquark in that of the antiproton).

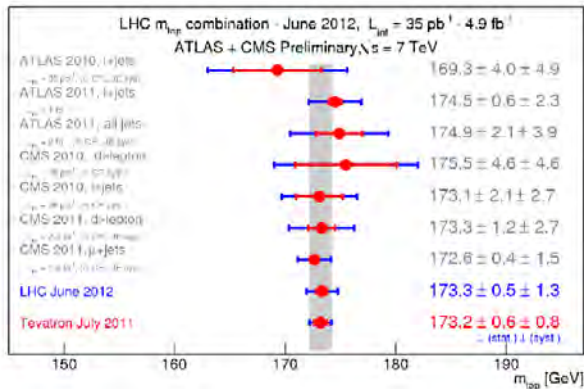
Although this is somewhat expected, the observed value is much larger than the NLO predictions. Some uncertainties involved in the calculations may be large but the effect may be also produced by some unknown effect, specially because the effect increases with the mass of the produced pair.

At the LHC the available energy and production yield motivates a more precise study of the effect. However, the symmetric initial state prevents the realization of exactly the same measurement. On the other hand, the matter-dominated initial state introduces differences in the rapidity distributions of the quark and antiquark that is related to the distribution studied at the Tevatron experiments.

The measurements of the asymmetry for the quantity  $\Delta|y| = |y_t| - |y_{\bar{t}}|$  performed by the two experiments [85, 86] show good agreement with the SM expectations. It should be remarked this does not exclude the Tevatron result, since there are no final model explaining the asymmetry. However, the LHC results exclude some proposed models and adds some additional information that is very useful for this subject, that is a good candidate to be one of the hot topics for the incoming years, specifically regarding top physics.

### – Study of $t\bar{t} + X$ production

Since the pair production cross section of top quarks is so large, it has become possible to start studying the properties of the top quark with the associated production of additional objects, usually radiated from the top. Sizes of the current datasamples do not allow detailed studies of the most interesting processes, as the production of a pair of tops and electroweak bosons, but current studies are showing the possibilities for the future running.



**Fig. 11:** Summary of the more relevant measurements of the top-quark mass at the LHC, including the combined from the two experiments and the comparison with the best Tevatron combination.

On the other hand, other processes that have not been studied in detail are already reachable for accurate comparison with the SM predictions. Two examples are given by the production of jets in association with a top pair [87] or even the production of bottom jets [88]. These measurements are in good agreement with expectations and are setting strong constraints on the model predictions in regions that were not investigated before.

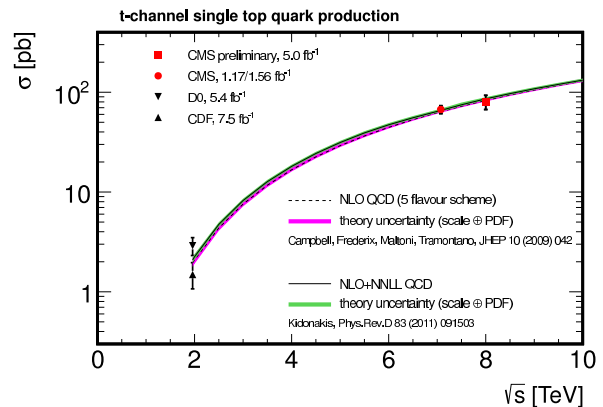
In summary, the LHC has been proven as a *top factory* allowing a high rate of produced top quarks to perform very detailed measurements of its properties. It is expected that the precision of these will increase with the future samples, providing information and constraints for models related to the less known of the quarks in the standard model. Therefore it is not an exaggeration to claim that particle physics has already entered in the era of precision in top-quark physics.

### 5.3 Single-top production

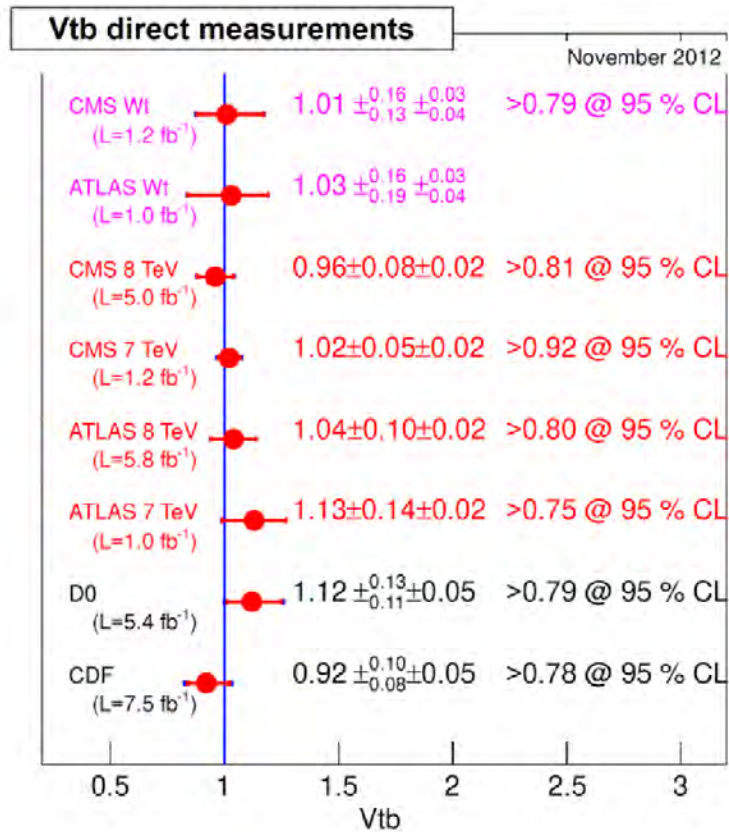
A very important topic regarding top production is that of *single top* that is dominated by electroweak production of top quarks. The process, observed at the Tevatron, has not being studied in detail until the arrival of the LHC, in which the available yields allow accurate comparison to the theory.

In the production of single top there are traditionally three channels under consideration: the t-channel (via a W exchange) which is the one with the highest cross section and sensitive to the bottom-quark content of the proton, the s-channel (via virtual W production) and  $Wt$  production, which was not observed at the Tevatron. From them, the t-channel is relatively easy to be studied at the LHC and current results have reached a good precision and even allowed separate studies of the quark and antiquark production. Figure 12 show the measurements at CMS at 7 TeV and 8 TeV [89] and comparison with Tevatron measurements. Similar studies has been produced by ATLAS, with similar reach and conclusions [90]. Additionally, results on the s-channel were able to set limits on the process that are around 5 times the SM predictions [91]. However, the current analysis does not include the full data available. With more data the results will become much more relevant. It should be noted that the s-channel is more sensitive to possible anomalous production of particles.

Regarding the third channel, the associated production of a W boson and a top quark, both experiments reached the level of evidence using the 7 TeV sample [92, 93]. The observed distributions are in agreement with the SM expectations, but more data is needed to perform accurate comparisons. The 8 TeV data should allow the observation and first precise measurements of this process, although the analysis is a bit challenging due to the harder conditions.



**Fig. 12:** Measurements of the single-top production cross section in the t-channel by CMS at 7 TeV and 8 TeV. For comparison, measurements at the Tevatron experiments are also shown.



**Fig. 13:** Summary of all the direct determinations of the CKM element  $V_{tb}$  at the Tevatron and LHC experiments from single-top production.

Once the production of single-top events has been established, the study of them allows to provide information about the electroweak couplings of the top quark, specifically due to the sensitivity of the production mechanism to the CKM element  $V_{tb}$  ruling the coupling between the top quark, the bottom quark and the W boson. Several determinations of this quantity have been performed at Tevatron and LHC, as summarized in Fig. 13.

In conclusion, studies of the single-top production are starting to reach a precision that will put the SM under test in the unexplored sector of electroweak physics with top quarks. Without doubt, this will also contribute in the next years to complete the picture we have of this quark as a key piece of the SM and its link to its possible extensions.

## 6 Results on heavy-ion collisions

Although the main goal of the LHC is to understand the interactions at the highest energies (or shortest distances), this collider also allows to produce extreme conditions in terms of energy density, pressure affecting baryonic matter. This is achieved by colliding heavy-ion nuclei, as it is the case of lead. The main goal is to try to study the strong interaction at lower levels, i.e. investigate concepts as confinement, thermal phenomena, chiral symmetry and so on, more closely related to the conditions affecting quarks and gluons in the early universe than the clean parton-parton collisions usually studied at the LHC when colliding protons.

Also in the case of the LHC the increase in energy represents a big step forward in studies of heavy-ion collisions: the experiments at RHIC were intended to discover the production of strongly-interacting perfect fluid. The LHC experiments shall characterize the details of this new class of matter

with the increased precision. For that, one of the most useful quantities is the *elliptic flow*, defined as the second momentum of the azimuthal distribution of produced particles. It contains very important physics information because larger values of the quantity indicates the presence of viscosity in the medium at the early times after the collision. Such values were observed at RHIC and by ALICE [94], confirming the expectations from hydrodynamic models. Additionally, ALICE has measured the elliptic flow and production yields (and ratios) for specific particles, as e.g. in [95] identified via its sophisticated detector subsystems. Some of the results are a bit unexpected, as the reduced production of baryons with respect to pions, which may be pointing to some presence of hadronic rescattering, an effect never observed. Other interesting measurements have already been performed by the collaborations with the aim of quantifying the characteristics of the collisions, as studies of higher-order harmonics (as in [96]), or particle correlations, and the studies related to the measurements sensitive to the Chiral Magnetic Effect [97] which is a fundamental study in the heavy-ion program at the LHC after the first hints at RHIC.

However, most of the current studies in heavy-ion collisions are more pointing to the confirmation of the results found at RHIC in order to tests new tools and fix a solid base to go beyond in terms of energy and sizes of data samples. In fact, it is in terms of hard probes of the created medium where the LHC experiments have clearly go beyond previous experiments.

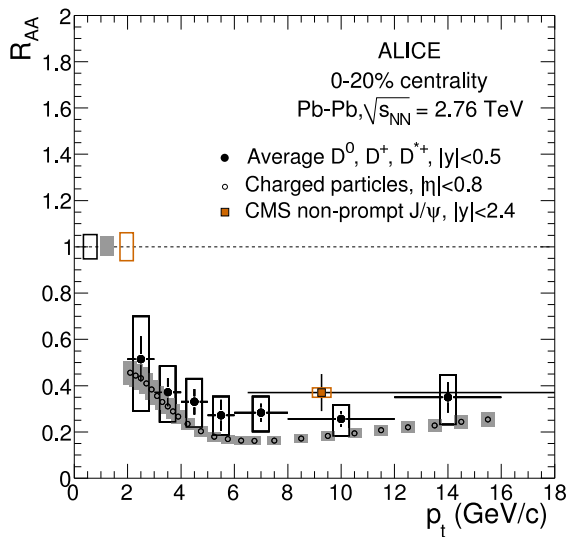
ATLAS was the first one presented a result on jet quenching [98], in which one expect dijet events produced from hard parton interactions in lead-lead collisions are observed as assymetric production of jets: opposite to a produced jet with large transverse momentum it is not straightforward to find a second jet, as in the usual proton-proton collisions. In fact a factor 2 of suppression in central collision is observed, very independent of the jet momentum. This is explained by the presence of a strongly interacting medium which affects more one hard parton than its companion, and therefore giving the impression of disappearance of jets.

In addition to jets, it has been very common the use of hard photons as probes of the medium. Photons are transparent to the medium, so they are perfect to quantify effects on jet quenching in the production of  $\gamma$ +jet, as in [99]. However, photons may also be coming from the hadrons in the medium, or in the final state, so they represent as small limitation that the LHC experiments may avoid with the use of more massive probes that were not available at RHIC: the weak bosons. Currently the experiments have been focusing on detecting the presence of those bosons, since available data samples does not allow its use as actual probes, e.g. in Z+jet production. However, the detection of leptonic Z bosons by CMS [100] and ATLAS [101] have already allowed the first differential measurements to characterize the production of these ideal probes, completely insensitive to initial state or hadronization and for which the medium is transparent. Studies of the W bosons have also been performed [102] and have already provided interesting confirmation regarding proton-neutron differences: isospin effect yields a reduced asymmetry in charge with respect to proton-proton collisions at the same energy per nucleon. Again larger samples are needed for more detailed studies, but the LHC is probing all its potential in heavy-ion collisions.

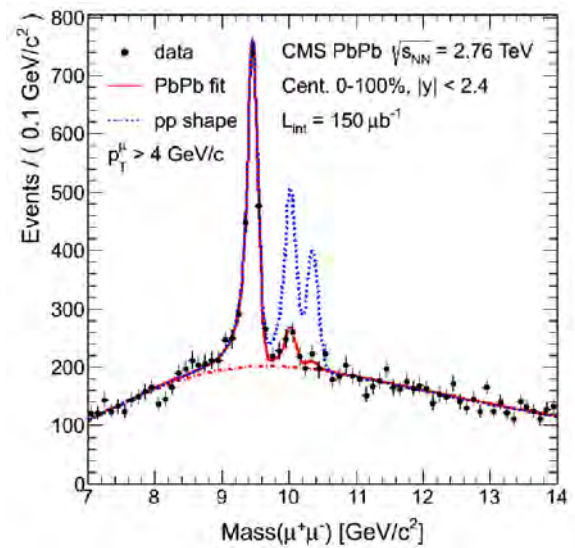
Another area in which the LHC allows to reach much further than RHIC is the study of heavy-flavour production. As in the case of proton-proton collisions, the possibility of identifying secondary vertices allows specific studies to be performed. In fact ALICE has shown its great capabilities with the reconstruction of open-charm mesons, D mesons [103] which are not only nicely observed but also used to perform measurements, like the one shown in Fig. 14, which probes the confirmation of suppression for open charm in central collisions, in good agreement with more inclusive studies. The aim of using open-charm mesons (and perhaps B mesons) is that they bring the possibility of quantifying differences in the energy loss in the medium between heavy or light quarks and even gluons.

But the identification of heavy-flavour states is much more powerful in the dilepton resonances, specifically for the quarkonia states. They have a long history of being studied in heavy-ion collisions due to their clean signature and the big theoretical/phenomenological knowledge on them. Regardless





**Fig. 14:** The nuclear modification factor with respect to proton-proton measured in lead-lead collisions for D mesons in the most central events as measured by ALICE. Data (black dots) are compared to the nuclear modification factors of charged particles (open circles) and non-prompt  $J/\psi$  from CMS (squares).



**Fig. 15:** Invariant mass of dimuon pairs measured by CMS in the region of the  $\Upsilon$  family as produced in heavy-ion collisions (dots and red-line fit). Comparison to the data from proton-proton collisions normalized to the  $\Upsilon(1S)$  peak (blue dashed line) shows the sequential suppression of the family in heavy-ion collisions.

of being colourless they are sensitive to the medium since they rely on the strong force to keep the two quarks bounded. In fact these states are affected by screening effect and they become an actual thermometer of the medium: the larger the radius of the system (larger for e.g. 2S states than 1S) the larger the screening. Therefore we expect to observe a *sequential suppression* or *melting* within the quarkonia families: less bound states are more suppressed than those that are more bound. This has been clearly observed in measurements by CMS [104] for the  $\Upsilon$  family, as shown in Fig. 15. Clearly the excited states are affected more in relative terms than the ground state when comparing results from lead-lead collisions with those of proton-proton at the same energy per nucleon. This is an additional confirmation that a strongly interacting medium is created in the relativistic heavy-ion collisions at the LHC.

It should be noted however that even if the qualitative picture seems clean, the quantitative details do not completely fit, so further measurements and theoretical developments will be needed in order to fully understand the generated medium. Such kind of studies are already in place, as the measurements of  $J/\psi$  suppression by CMS [105] (in central rapidities) and ALICE [106] (in forward rapidities), probing the nice complementarity between experiments. However the agreement in the suppression does not apply to the observation by CMS that  $\psi(2S)$  is less suppressed than the  $J/\psi$  for transverse momenta larger than 3 GeV, something not confirmed by the ALICE measurements.

In conclusion the heavy-ion program of the LHC experiments is already providing interesting results bringing the field to unexplored areas with a new energy regime and new possibilities, like the use of new available tools and probes. The prospects for the future, with further analyses of the data, including the  $30 \text{ nb}^{-1}$  collected for proton-lead collisions (as the previews in [107, 108]), will help towards the ultimate goal of the program: detailed characterization of QCD thermal matter by means of precise measurements from heavy-ion collisions at LHC.

## 7 Searches for the SM Higgs boson

The SM structure and its implications in the description of the Universe is based on the presence of a field, known as *Higgs field* that is responsible for the symmetry breaking giving rise to the electromagnetic and weak interaction and also to give masses to the weak bosons. In this process, a single degree of freedom is translated into a scalar particle, *the Higgs boson*, that should be observed and whose coupling to the fermions are introduced in such a way that these last ones acquire the masses that are forbidden by the symmetry before it gets broken.

This particle is therefore the missing keystone of the SM and it was extensively searched for in previous colliders without success. The good performance of the SM strongly motivated the existence of the particle, and the measurements and fits from pre-LHC colliders pointed to a mass of around 100 GeV [109].

Under this situation, the LHC started collecting the data that should provide light to the existence of this boson and eventually find it. This was the most important search for the first years of the LHC experiments and for this reason it deserves a full section describing the analyses and the strategy to follow in order to observe the presence of the boson and also the related measurements which are aiming to confirm whether the observed resonance actually matches the properties expected for the SM Higgs boson.

### 7.1 Strategy to search for the boson at the LHC

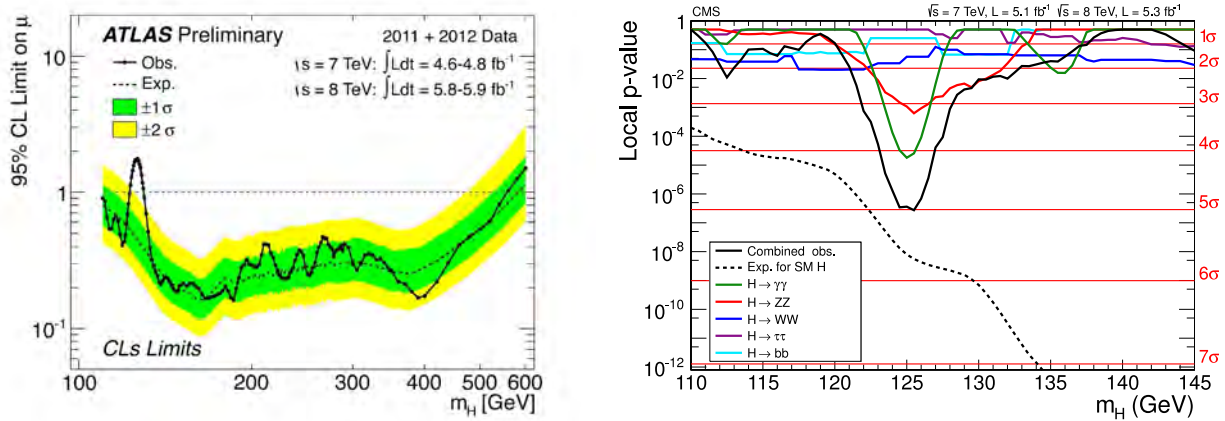
Before the LHC had collected enough data for being competitive in searches of the Higgs boson, the results from LEP and the Tevatron were the richest source of information. In fact, LEP had excluded at 95% C.L. the SM Higgs boson below 114 GeV and its measurements had constrained the mass of the Higgs to be around 100 GeV.

In the case of Tevatron, the direct searches were excluding a Higgs around 165 GeV, leaving the available regions to be clearly separated into two: The low-mass region, for masses between 115 and 160 GeV, that was very strongly motivated. The second region, with relatively high masses beyond 170 GeV, was less motivated, but still not discarded, specially considering that the motivation was assuming negligible effects from possible BSM physics (or more complex Higgs models).

The first step therefore for the LHC was to look into these two regions and during 2011 all channels were considered to investigate all the mass ranges. For low masses, although the decay is dominated by that to bottom quarks, the involved channels were those having the Higgs decaying into ZZ (in 4 leptons) or  $\gamma\gamma$ , with some information from the WW,  $\tau^+\tau^-$  and  $b\bar{b}$  decays in all accessible production modes. For high masses the most useful channels were those involving decays into WW and ZZ in all possible signatures. With this approach the two experiments presented results on December 13<sup>th</sup> 2011 with the data collected at 7 TeV. The results presented at that time led to a complete exclusion of the Higgs boson in the high-mass region (up to more than 400-500 GeV) and most of the low-mass one, leaving alone a small window around 125 GeV.

In that window the exclusion was not possible because both experiments saw an excess, not completely significant but enough to prevent exclusion of the presence of a SM Higgs boson. The excess was appearing in several of the channels. Naturally, the presence of a resonance in the most motivated channels to detect the SM Higgs boson was a clear suggestion that such boson was the responsible for the excess, so all the focus from that moment was to intensively search for a possible boson with a mass around 125 GeV whose properties were close to those expected for the SM Higgs boson.

This effort was designed to be applied to the 8 TeV data collected right after the Winter in 2012 and the idea was to maximize sensitivity in the two most sensitive channels at that mass (4-lepton ZZ and  $\gamma\gamma$ ) and also look at the complementary channels (WW,  $\tau^+\tau^-$  and  $b\bar{b}$ ) that could provide some further sensitivity and also some additional information regarding the nature of the boson: more couplings involved.



**Fig. 16:** On the left, 95% C.L. limit on the ratio of the cross section over the SM expectation for the production of a Higgs boson as a function of its mass as obtained by ATLAS at the time of ICHEP-2012. Observed limit is compared with the expected limit (in absence of such a particle) and the uncertainty intervals. On the right, the local p-values for a similar study in several analyses by CMS. In both cases a very significant excess is observed around 125 GeV that is interpreted as observation of a new particle, likely the SM Higgs boson.

In parallel more analyses were still considered in order to complete the pictures, even those that were looking (and excluding) the presence of a SM Higgs boson at higher and higher masses.

All these analyses are described in the following sections.

### 7.2 Analyses for the discovery (ICHEP-2012 results and afterwards)

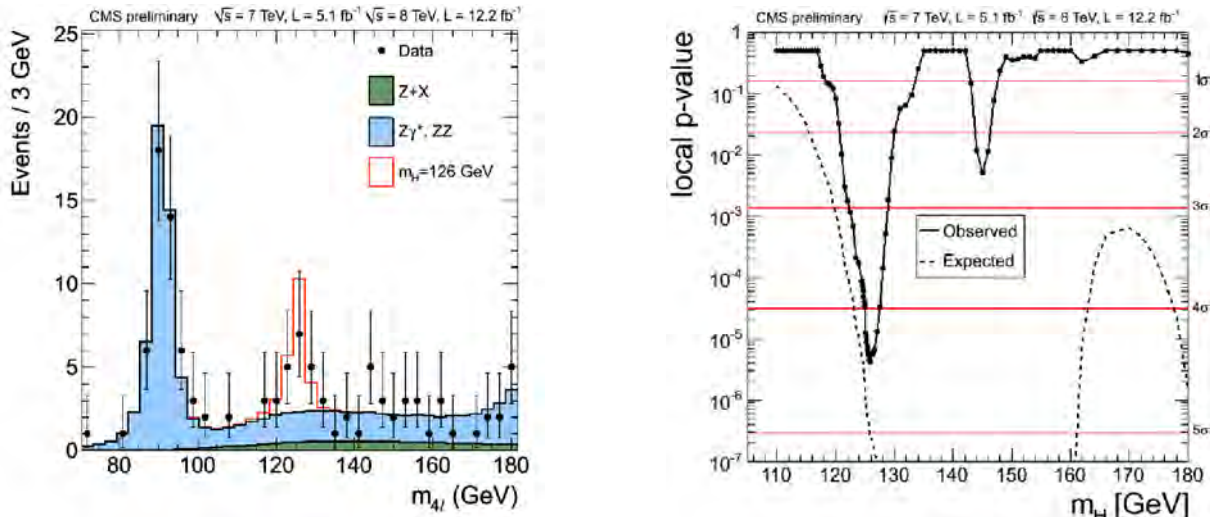
At the time of ICHEP-2012 the size of the available data at 8 TeV was comparable to that collected at 7 TeV, allowing already enough sensitivity to perform statements on the boson. Both collaborations presented results in the main channels on July<sup>th</sup> 2012, and they confirmed the presence of a new boson at the discovery ( $5\sigma$ ) level. The presented results are summarized by the plots in Fig. 16, where the results from the statistical analyses of the studies are shown.

The measurements performed at 8 TeV also increased the precision on the knowledge of the boson and in general tend to confirm its nature as that of the SM Higgs boson. Later improvements to the analyses and the addition of the data that was provided by the LHC during 2012 have brought additional support for this hypothesis. However, some questions are still to be investigated and further data would allow more precise measurements in the future. Here we will discuss some of the more relevant results bringing to the current knowledge about the boson discovered at a mass of 125 GeV.

In the case of CMS, the  $H \rightarrow \gamma\gamma$  search [110] is performed by using several categories of diphoton (for inclusive production mode) and two categories for tagging Vector-Boson Fusion (VBF) processes. It should be noted that VBF is very important because it is sizable (mostly because the leading Higgs production occurs via loops) and it involves different couplings than the dominant mechanism, e.g. it is very important for fermiophobic models.

With all those categories, the analysis is able to achieve a significant excess of  $4.1\sigma$  with a yield a bit higher than expectation.

In addition to that, the 4-lepton search was dealt in this collaboration with the use of a kinematic discriminant that accounts for the fact that the Higgs boson is a scalar. This kind of tools have made that this analysis [111] is the central reference for measuring the properties of the boson, as described below. As shown in Fig. 17 the channel has very little background and the signal is clearly observed in spite of the low yield. The significance of the excess at a mass of 126 GeV is very high, although in this case the yield comes a bit lower than the SM expectation, but still in agreement.



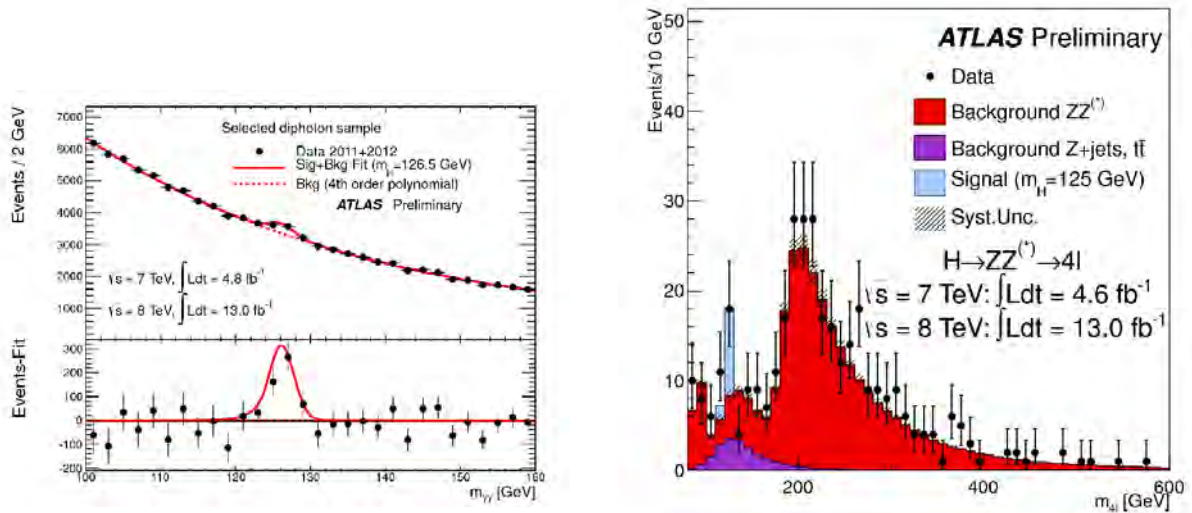
**Fig. 17:** On the left, the mass distribution of 4 leptons in events selected for the Higgs search at CMS. Data (dots) are compared to the background expectation (solid histograms) and a Higgs signal with  $m(H)=126$  GeV (red line). On the right, local p-values associated to the same analysis, with a clear excess for a Higgs mass around 125 GeV.

In addition with the most sensitive channels, CMS has put a lot of effort on the secondary channels which are giving additional constraints about the boson, with a small sensitivity. Specifically, the  $WW$  decay also suggests the existence of a boson, but with a yield on the lower side [112]. The  $\tau^+\tau^-$  shows clear limitations on the size of the data sample and although the result is compatible with a SM Higgs, it is also in agreement with the background-only hypothesis [113]. A similar conclusion is extracted from the decay into bottom quarks [114], in which the Higgs need to be observed in the production associated with a weak boson, in order to keep the dijet background under reasonable limits. The studies of diboson production described in section 3.4, specifically in the semileptonic channels, provide a solid support to the search of the boson in this decay channel. In any case, more data will provide stronger constraints on the fermionic decay channels, currently compatible with the existence of the SM Higgs boson but with small significance.

From the ATLAS side, also several updates came after ICHEP-2012, bringing further confirmation to the signal and, as in the CMS case, higher precision in the results. The diphoton search [115], performed with several categories, has lead to a very strong signal, which approaches the level of being very high when compared to the SM expectation with a signal strength value approaching a factor of 2 (being 1 the SM prediction). Dedicated studies of this value in a per-channel basis does not indicate anything striking, but uncertainties in those cases are large since it is the combination of them which is bringing the high significance of the signal. Plot on the left of Fig. 18 shows the invariant mass distribution of diphotons in which the resonance at a mass around 125 GeV is clearly observed.

As in the diphoton search, the 4-lepton channel in ATLAS gives a signal strength higher than the expectation, although in this case in agreement with the SM value (and with the CMS result). The study of this final state [116] is performed by exploiting the kinematical properties of the decay products from a spin-0 particle. As shown in the plot on the right of Fig. 18, the signal is clearly observed with a reasonable amount of background, which leads to this channel as the main reference to measure the properties of the boson, as in the case of CMS.

Regarding the complementary channels, ATLAS also puts a big effort on those with similar conclusions to those obtained by CMS. In the case of the decay into bottom quarks [117], sensitivity has not yet reached the level to allow quantitative statements about the boson to be made. The other two channels [118, 119] give higher yields than expected, but still with large uncertainties. In the case of  $\tau^+\tau^-$ ,



**Fig. 18:** On the left, invariant mass of two photons in the search for the Higgs decaying into diphotons performed by ATLAS. The fitted background is subtracted in the plot below in order to enhance a resonant excess close to 125 GeV. On the right, invariant mass distribution of 4 leptons in events selected for the Higgs search at ATLAS. Data are compared to the background and a signal hypothesis with  $m(H)=125$  GeV.

the value seems to be high in the case of the main production channel, but in VBF and in associated production with a weak boson (VH) the signal strength is clearly on the low side [119]. It is too early to be considered a problem since the uncertainty is still large enough to cover the SM value within  $1\sigma$ .

### 7.3 Post-discovery goals: measuring the properties

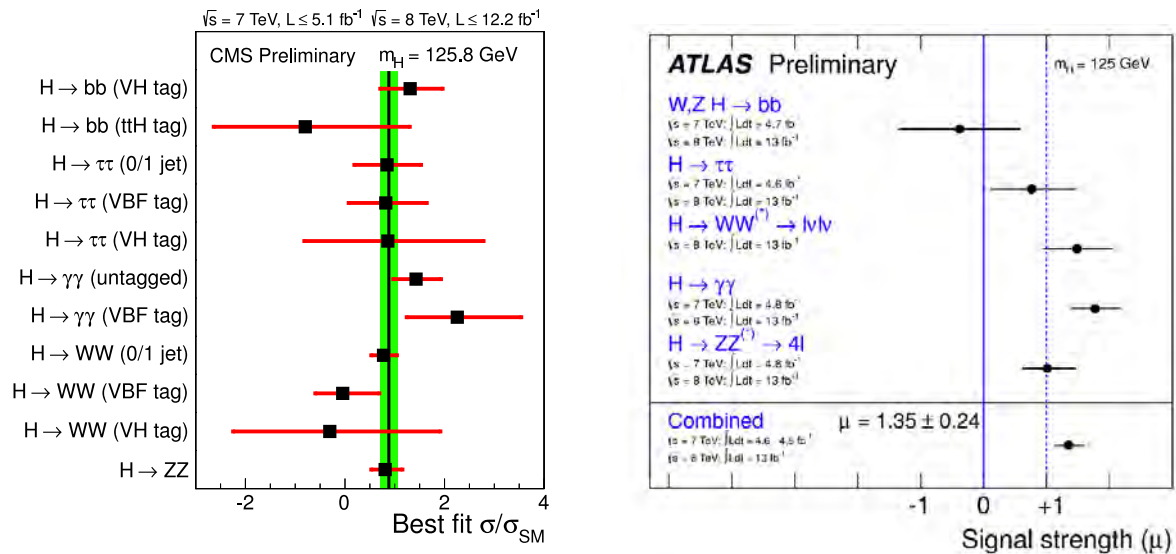
As described in the previous section, a new boson has been observed and its properties are compatible to those expected from the Higgs boson of the SM. With the additional analysis the picture is getting more complete, but precision needs to be improved to extract further conclusions.

One of the goals in the incoming *post-discovery* years is the measurements of all the properties. This has been already started, and some answers are already provided, as we will discuss here.

The first set of results is the comparison of the signal strength for the several channels that have been investigated. The results are summarized in the plots of Fig. 19. As mentioned in the previous section, values are not completely matching the expectations from the SM, but they are not significantly discrepant. More data will be needed to reduce the uncertainty and investigate possible anomalies in the production and decay mechanisms. Explicit disentangling of the couplings show they are fully compatible with the SM expectations, as in [120].

After the production mechanism has been checked, the first obvious property to measure is the mass of the found resonance. Dedicated studies has been performed at the two collaborations using the most sensitive channels. In the case of CMS, the last study has been based on the 4-lepton sample and provides a mass value of  $m(H) = 126.2 \pm 0.6(\text{stat}) \pm 0.2(\text{syst})$  GeV [121]. In the same analysis, studies of the spin and the parity leads to the conclusion that the data clearly favours a pure scalar versus a pseudoscalar. Additionally, data is not precise enough to distinguish between spin-0 and spin-2 particles in this channel.

In the case of ATLAS, the results presented in [122] show some tension between the masses extracted from the 4-lepton and the diphoton channels. In the first case a value of  $m(H) = 123.5 \pm 0.9(\text{stat}) \pm 0.3(\text{syst})$  GeV is obtained. For the second, the value is  $m(H) = 126.6 \pm 0.3(\text{stat}) \pm$



**Fig. 19:** Signal strength in the several channels sensitive to a Higgs boson with a mass close to 125 GeV in CMS (on the left) and ATLAS (on the right). SM prediction should be centered at 1, which is compatible with the measured values and with the combined average.

0.7(syst) GeV, in better agreement with the measured value at CMS using the 4-lepton channel. This discrepancy will require some further investigation and perhaps data to be understood. It should be added to the issue that the signal strength values as measured by ATLAS tend to be higher than the SM expectations.

In addition to the mass measurement, studies of the spin and parity has also been performed by ATLAS [123]. They are similar to those by CMS, but more complete since information is also extracted from the  $H \rightarrow \gamma\gamma$  analysis. This has allow to add more sensitivity to the distinction between spin-0 and spin-2 particles.

#### 7.4 Other searches for SM-like Higgs and within models of new physics

Even though a boson that is a good candidate to be the Higgs as predicted by the SM has been found, other analyses looking for SM-like Higgs bosons are still of interest. The main motivation is that they may be sensitive to scalar resonances with a mass larger than that of the boson, or smaller but with lower production cross sections.

Most of these searches are following very closely the searches for the SM Higgs at the correspondent masses, since they inherit from analyses performed before the boson was observed. They are naturally diverging from the optimal search for the SM Higgs, in order to look for similar particles, but not with exactly those properties of the SM Higgs. Many searches has been performed by ATLAS [124] and CMS [125] and have computed limits for possible presence of particles that are SM-Higgs alike, since no hint for a resonant scalar has been seen.

Furthermore, several BSM theories include the modification of the Higgs-sector, which implies that other Higgs particles may be present in Nature, even with the presence of the SM one. The suggested discrepancies in the Higgs properties add further motivations for this kind of models. Note we discussed them here even if searches for BSM physics are included in sections 8 and 9.

As usual in searches for new physics, supersymmetric models are the most attractive to be considered. In the case of Higgses, Supersymmetry (SUSY) requires the presence of at least five Higgses, one basically like that predicted in the SM and others that are relevant due to their properties: charged Higgses and Higgses with enhanced couplings to bottom quarks and  $\tau$  leptons. This later case motivated

the search for a Higgs decaying into  $\tau^+\tau^-$  interpreted in SUSY models. Lack of any observed signal brings the experiments to use the results [126, 127] to set constraints in the SUSY parameter space.

In addition, searches for charged Higgs have been performed in order to look for their presence in decays of the top quark. CMS has focused on the  $\tau$  channel [128], looking for an anomalous presence of  $\tau$ -based decays with respect to other leptonic channels. Limits were set for several models due to the good agreement of the data with the W-only-decay hypothesis. In the case of ATLAS, one of the investigated channels was  $H^\pm \rightarrow cs$  [129], in which the presence of a dijet resonance not peaking at the mass of the W boson will be identified as a signal. In addition, we expect a lower yield due to the competing channel that is purely hadronic (assuming that the charged Higgs decay preferably into that channel). Data does not confirm these expected anomalies, so additional limits are set for this kind of model.

Aside for the basic SUSY models, other extensions of the SM incorporate modifications of the Higgs sector and therefore they have been searched for. There are many possibilities here, and several classes of Higgses show up. However, we should emphasize that some of them yield topologies that may have been missed due to kinematic selection, as it is the case of Higgses with low masses (as the dimuon resonance search in [130]) which may be produced just as boosted objects due to their own couplings. Other possible exotic particle in the Higgs sector is the presence of doubly-charged particles whose searches, as the one in [131], have not reported any visible discrepancy with respect to the expected SM backgrounds.

In conclusion, no significant hint of alternative or extended Higgs sectors has been found to complement the boson observed at a mass around 125 GeV. However, this does not imply that the physics beyond the SM is out of reach, since the Higgs sector is well known for providing very elusive particles. For this reason, searches of new particles have been performed independently of the discovery of the possible Higgs, as discussed in the following sections.

## 8 Searches for new physics

As it has been discussed before, the LHC is intended as a machine to bring information about new physics beyond the SM. The possibility that the Higgs boson has been found does not only confirm the validity of the SM, but also its limitations that should be investigated to find even more correct answers about the structure of the Universe at the smallest distances.

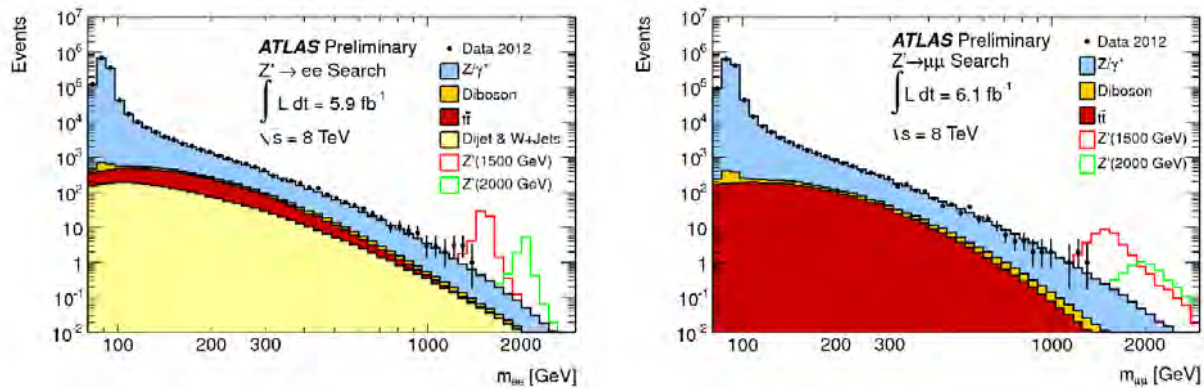
Finding these answers at the LHC requires a huge effort in order to cover the many possibilities, and therefore corners of the parameter space. This makes the search topic a very broad field of investigation. In this report we just summarize the most interesting searches of all those developed at the LHC.

Within the searches for BSM physics, the models involving SUSY are strongly motivated due to their good theoretical performance to solve the SM limitations. Specifically the doubling of the particle spectrum, in order to have a supersymmetric partner to each SM particle, allows a very rich phenomenology that translates into many analyses investigating several types of final state topologies. Those are discussed in section 9.

On the other hand, there are well-defined alternatives to supersymmetric models that also provide possible explanations to the issues of the SM as the full description of the Universe. In the following subsections we focus on summarizing the searches for these alternative models.

### 8.1 Searches for unknown high-mass resonances

When looking for new physics, the more direct approach is to look for particles that are not included in the SM spectrum. For that, the search for resonances decaying into detectable and well-known particles is the simplest approach. Some of these resonances are naturally predicted in extensions of the SM, specially with the addition of new interactions. Figure 20 shows the invariant mass of dileptons as measured by



**Fig. 20:** Invariant mass distributions of electron-positron (left) and dimuon (right) in dilepton events collected by ATLAS. Data (dots) are compared with the SM predictions (solid histograms) and some expectations for BSM resonances (lines).

ATLAS when looking for a massive resonance that would appear as a peak in those distributions. The comparison of the data with the SM expectation show very good agreement and results [132] (and [133] for CMS) are used to set limits on the production cross section for resonances, and lower mass limits on possible Z-like particles in the order of 2.5-3 TeV.

Similar to the lepton search, the production of dijet resonances has also been considered, as in the CMS result documented in [134], in which special treatment has been performed in order to separate between resonances decaying into gluons or into quarks (or a mix). Also in this case, a good agreement has been observed, but the main issue is how to handle the huge background at the lower invariant mass, that forces to reject events even at the trigger level.

This has been the testing analysis of a new technique, called *data scouting*, which allows to collect interesting events passing around the trigger limits. The idea is to collect events at a higher rate but storing only the final reconstructed objects, which allows the reduction of the data content per event. This permitted CMS to trigger and perform studies for lower invariant masses with competitive results [135] even with a reduced datasample of  $0.13 \text{ fb}^{-1}$ .

When looking for resonances, the presence of neutrinos is not a limitation, and the search is also extended to the use of the *transverse mass* of a lepton and the  $E_T^{\text{miss}}$ , defined as

$$M_T = \sqrt{2 \cdot p_{T,\ell} \cdot E_T^{\text{miss}} \cdot (1 - \cos \Delta\phi_{\ell,\nu})}$$

to investigate the presence of new resonances decaying into a charged lepton and a neutrino. In the case of a resonance, this variable shows a Jacobian peak that is on top of a smooth background. The current results, as those in [136], do not show any hint of such type of structure, and limits on production of W-like particles has been set.

However, when we talk about limits on very massive W-like particle, a possible decay channel is into a top and a bottom quark, which is not allowed for the W. This was investigated by ATLAS [137] and found no sign of a resonance decaying into those quarks, and independently of the number of the identified b-jets. It should be noted that the searches of this kind of resonance have become very powerful at the LHC due to the available energy for producing high-mass resonances decaying to the most massive particles in the SM spectrum. This is also confirmed in the study of resonances decaying to weak bosons, which are predicted to appear in several BSM theories. A result by ATLAS has taken advantage of the trilepton final state to look for resonances decaying into WZ [138], providing a very competitive result, although usually this kind of search is performed with semileptonic or fully hadronic channels to make use of the larger branching ratio.



In fact, the energy at the LHC is so large and the possibility for producing resonance so large, that very massive object could appear and the decay products will be boosted, which may lead to dijet (e.g. from W) merged into one reconstructed jet. This has been turned into a benefit to enhance signal, by using merged jets to tag the presence of hadronically decayed bosons. The result of the analysis by CMS [139] shows the good performance of the boosted-jet tools. Unfortunately no sign of new physics was found.

A similar analysis by ATLAS looking for a resonance decaying into ZZ in the semileptonic channel [140] also exploits the merged jet topology to increase acceptance to very massive resonances and set a much constraining limit than that accessible by the obvious dijet topology.

Among the searches for resonances indicating BSM physics, one common topic is the studies of possible excited states of fundamental particles, which could be related to new physics (e.g. contact interactions or internal substructure). This is the case for the search of excited muon states decaying into a muon and a photon as the one by ATLAS [141] looking for the Drell-Yan production of a muon and an excited muon. The results are in good agreement with SM predictions for the most discriminant variable: the invariant mass of the two muons and the photon, which allows to set stringent limits in the possible scale for such a excited state to exist.

In addition to the searches for resonant states in the two-body decays, the high masses accesibles at the LHC allows the searches for more complicated topologies, with more objects in the final state. One example is the search for boosted resonances decaying into three jets. The search performed by CMS [142] assumes pair-production of these objects, and therefore the idea is to study three-jet ensembles whose transverse momentum is large but the corresponding mass may show a peak structure related to a decaying resonance. The requirement of large transverse momentum allows the reduction of the combinatorial background, for which the mass and the transverse momentum will show a correlation. Although the result of the analysis does not show hints of any possible resonance, the used technique can be used in other searches in the future. In the current case, limits are set on the existence of resonances.

Another alternative that is open at the LHC is the cascade decay with initial massive objects sequentially decaying into states. A very symmetric case considered at CMS consists on the pair production of objects (e.g. technicolour particles) decaying into pairs of particles (e.g. other lighter state in the technicolour spectrum) which decay into dijet. This process will lead to an 8-jet topology in which there are resonant peaks in four dijet masses, two 4-jet masses and perhaps in the 8-jet mass in case the original pair-production occurs from the decay of a single-produced particle. All this information is combined into an artificial Neural-Network to enhance signal-like topologies. The results [143] show that there is no peak structure on top of the combinatorial background coming from usually-produced 8-jet events and limits has been set for models motivating this kind of signature.

## 8.2 Searches for leptoquarks

One special case of pair-produced resonances that are motivated by unification models is *leptoquarks*, particles having both lepton and baryon numbers. They are detected via their decay into a lepton and a quark, which gives a resonant peak in the invariant mass (in the case of charged leptons) or significant excess in  $E_T^{\text{miss}}$ -related variables (in the case of neutrinos).

Since these particles carry colour, they are pair-produced with a large cross-section, giving rise to clean signatures due to the leptons in the decay. Furthermore, they also have a rich phenomenology, since these particles could be of different classes (scalar, vector) and also appear in different generations, although they are usually not mixing fermions of different families.

The basic analyses, mostly oriented to the first two generations are easily identified by the kind of lepton, which determines the generation we are focusing. Searches by ATLAS [144] show good agreement with the SM expectations for the  $eejj$  and  $\mu\mu jj$  final states. These results are used to set limits that are going beyond previous searches of these particles.

Since the first generations are not providing hints of leptoquarks, even in the channels with neutrinos, searches have also been focused on the third generation, where  $\tau$  leptons and bottom quarks are expected. Specifically the search by CMS [145] with the use of b-jets exploits the sensitivity given by the scalar sum of the transverse momenta of the decay products. The results are in good agreement with the SM expectations, and they are used to set limits on the leptoquark production, but also on the production of scalar tops within R-parity ( $R_P$ )-Violating SUSY models (see details in section 9.3), giving an explicit proof that searches of new physics are usually sensitive to several classes of models bringing similar final states, an in similar areas of the phase space.

### 8.3 Extradimensions and graviton searches

The extensions of the SM do not only consider the extension of the particle spectrum or the interaction sector. Several models introduce the modification of the structure of the Universe by incorporating additional dimensions, that would be microscopic and whose existence may explain the large scale difference between the electroweak interaction and gravitation. The idea is that the new dimensions will be forbidden to the SM particles and effects, while gravity expands in all the available dimensions. The signatures will be striking with the production of gravitons (producing large  $E_T^{\text{miss}}$  since they escape detection) and SM particles, leading to single-photon (monophoton) or single-jet (monojet) topologies,

These have been looked for by the collaborations. As an example, ATLAS has looked for events with a photon with large transverse momentum that is accompanied with large  $E_T^{\text{miss}}$ , which is the most significant variable to identify the presence of new physics [146]. Good agreement is observed with respect to the SM expectations for this signature, dominated by undetected weak bosons (neutrino decays) in association with a photon. Also some background contribution is present due to detector effects generating artificial kinematics looking like the signal.

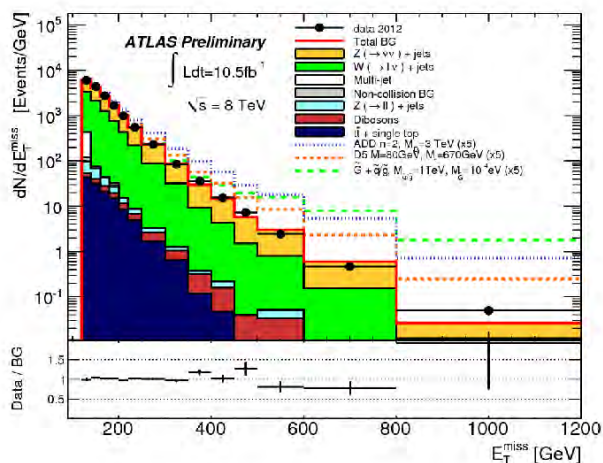
Furthermore, ATLAS and CMS have also looked for the monojet topology [147, 148]. Although the main motivation for this signature is the production of gravitons produced in association with quarks, there has been an increase use of this kind of search for studying the production of invisible particles (as generic Dark Matter candidates) in a model-independent way, being the jet balancing the  $E_T^{\text{miss}}$  produced by initial-state radiation. This keeps a small fraction of the total signal, but allows to look for hard-to-detect particles that may be copiously produced at the LHC collisions. It should be remarked that this makes a strong case when compared to the more clean monophoton signature: results are more sensitive to other classes of models.

The results of the monojet searches has also found good agreement with the SM predictions. Figure 21 shows the  $E_T^{\text{miss}}$  distribution of the ATLAS analysis [147], that has also been used to set limits in the production of gravitino from the decays of squarks and gluinos.

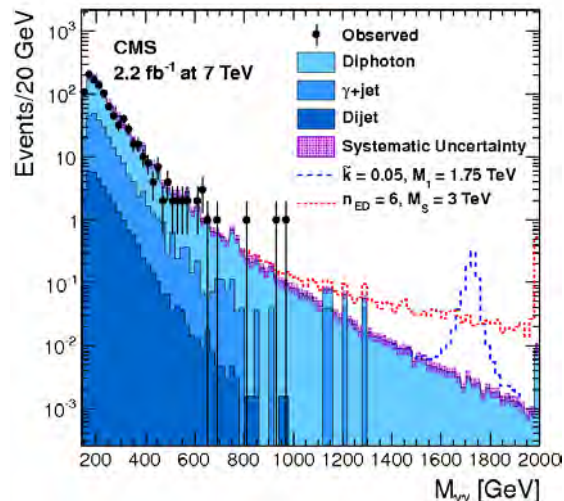
Another possibility related to extra-dimensions and accessible production of graviton is that particles may appear as Kaluza-Klein towers which sequentially decay into less massive objects. Specifically, gravitons may appear as diphoton resonances, which is an easy-to-identify signature, but it suffers from large backgrounds. Anyway, they have been investigated by the LHC experiments, as the analysis in [149], and no hint of such a resonance has been found on top of the diphoton high-mass spectrum, as shown in Fig. 22, which also includes the expectation from a resonance as those predicted by Randall-Sundrum models and the expected effect due to a more generic model including additional dimensions.

### 8.4 New physics in the top sector and new generations

As discussed before, the top quark is usually suggested as the primary candidate to open the path towards new physics. Its large mass and coupling to the Higgs, which are the basic quantities related to the loose ends of the SM, make this quark a very attractive place to search for discrepancies with respect to the SM expectations, Since the first step to fix the hierarchy problem is to have a partner canceling the top-induced corrections to the Higgs mass, such a partner should be at reach of the LHC independently of its



**Fig. 21:** Distribution of the  $E_T^{\text{miss}}$  variable in events with a single jet with large transverse momentum. Data (dots) are compared to the background expectations (filled histograms) and possible signals for BSM physics (coloured lines).



**Fig. 22:** Distribution of the diphoton invariant mass in events with two photons with large transverse momentum. Data (dots) are compared to the background expectations (filled histograms) and possible signals for BSM physics (coloured lines) containing a Randall-Sundrum resonance or a generic extradimension model.

nature. Although the most obvious choice is a SUSY partner (see section 9.2), alternative options have been made, including the possibility of the existence of a very massive 4<sup>th</sup> generation.

One option considered by ATLAS [150] is the search for the pair-production of a top partner having an electric charge of 5/3 (of that of the positron). Appearing in several models, the decay into a top quark and a W boson allows to have good acceptance with same-sign dileptons and also to use the hardness of the event (scalar sum of transverse momenta of final objects) as discriminating variable. No significant discrepancy has been observed with respect to the low expected SM background.

As a general rule, the existence of additional generations (containing canonical or exotic particles) that would contain coloured particles more massive than the SM ones leads to very busy final states in terms of multiplicity and of energy. This is used in the optimization looking for this kind of topologies, being very common the requirement of hard events or with rare combination of objects (same-sign leptons, leptons and b-jets in high multiplicities and similar requirements). The performed analyses searching for a 4<sup>th</sup> generation, as those in [151, 152]. All searches have brought the conclusion that there are no hints for the existence of a 4<sup>th</sup> generation (in the reachable masses) nor of any new physics that may look like massive particles regarding busy final-state topologies.

### 8.5 Searches for very exotic signatures

The lack of success to find hint of straightforward BSM physics has open the possibility that Nature is not as predictable as we might think and the new physics may appear in some even more exotic signatures than those considered for the theoretically-motivated BSM final states. This has led to the study of final states that could have escaped the more traditional selection or based on models less related to the confirmed SM predictions, which bring to new classes of final states.

One option that has been considered is the production of microscopic black holes at the LHC collisions. Some generic properties of them from quantum gravity provide general rules of final-state expectations: high multiplicities and democratic treatment of objects. The search for this kind of events [153] was performed by exploiting that the scalar sum of transverse energies for the SM background presents

a shape that is independent of the multiplicity. Therefore the lower multiplicity events are used to get the shape that is compared to the data with high multiplicities. Good agreement has been observed and limits in a model-independent approach are set.

Other rare topology that is not commonly considered is the presence of long-lived particles that may escape even the trigger selection. Some of these particles appear in several models as quasi-stable particles. In this context, searches for charged massive particles (CHAMPs) by CMS [154] or more dedicated searches like the stable chargino using track-disappearance by ATLAS [155] are good examples of the possibilities beyond the usual approaches and how the detectors are used with non-standard event reconstructions to look for unexpected classes of particles. In this, we should also mention the search for magnetic monopoles (as that by ATLAS in [156]), whose existence is very strongly motivated due to the electric charge quantization and as part of the electromagnetic unification. The need for specific reconstruction of the events (since these particles are not electric charges, and behave very differently inside magnetic fields) add some complication to the analysis, but still the results are very competitive when compared to direct searches because of the possibility to produce them with high cross section at the LHC. In any case no hint for production of monopoles has been observed and further data will help to increase the sensitivity, specially with the addition of the dedicated experiment for this (MoEDAL [8]).

In conclusion, after the first datasamples provided by the LHC collisions have been analyzed, no discrepancy with the SM prediction has been found that could be considered as a significant hint of new physics or particles beyond the SM spectrum. The future running of the LHC at a higher energy and higher luminosities, discussed in section 10, should provide more information on the possible BSM physics.

## 9 Searches for supersymmetry

In SUSY models the particle spectrum is at least doubled [157], bringing a lot of possible processes that could distort the measured values with respect to the SM expectations. Depending on the considered process, the final state to be investigated is different, providing a rich phenomenology.

However, since at the LHC the initial state is based on partons, the dominant production mechanism is usually the production of coloured superpartners. In usual models they are produced in pairs since R-parity ( $R_P$ , a quantity being 1 for particles and -1 for superpartners) is conserved. In addition, the conservation of  $R_P$  implies that the lightest SUSY particle (LSP) is stable and a Dark Matter candidate.

These basic properties allow to make general analysis in searches for SUSY which focus on specific parts of the spectrum. In addition, this also brought a new way of interpreting the results which are based on “simplified models” which provide well-determined processes for the given final states. This has simplified the interpretation of the results in terms of the possible theoretical models. On the other hand, the more traditional, “full model”, approach are still advantageous to interpret results from different analysis and experiments within a common framework.

Independently of the model the most basic search for SUSY is to look for jets and  $E_T^{\text{miss}}$ . The latter being a hint of the stable LSP, and the jets appearing as the decay products of coloured superpartners, which are the ones associated with larger production cross sections: squarks and gluinos. These analyses are just dependent on the reconstruction of the  $E_T^{\text{miss}}$  and they try to quantify its presence with variables that are less sensitive to misreconstruction. In addition several categories are investigated in order to be sensitive to different kind of SUSY processes. The categories are usually identified by the hardness of the event (with  $E_T^{\text{miss}}$  or momenta of jets), the multiplicity of jets, or the multiplicity of b-jets.

The analyses by the collaborations, as those in [158, 159], do not show any significant discrepancy with respect to the expected backgrounds. Results are used to set limits in several types of models, and are typically excluding the presence of squarks (of the first generations) and gluinos below 1-1.5 TeV.

In the case of massive squarks, it is feasible to produce gauginos that are lighter but still hard to produce directly from the proton collisions. These gauginos may decay in leptons with large transverse

momenta which simplify the identification of the events at the trigger and reconstruction levels. Both collaborations have searched for SUSY events in final states with leptons, jets and significant  $E_T^{\text{miss}}$  [160, 161] and the results shows good agreement with the SM expectation. Results have been used to set limits on the production of SUSY particles that produce leptons in the final state. It should be noted that the studies with leptons include the  $\tau$  lepton (as in [162]) since they provide increased sensitivity to the case of Higgsino-like gauginos.

When one consider leptons in the final state, the presence of multileptons may be a good hint of SUSY due to the reduced SM backgrounds. Specially when there are at least three leptons and significant  $E_T^{\text{miss}}$ , which is the golden final state detecting the production of a pair of chargino and neutralinos decaying leptonically or even production of scalar leptons. The background of these kind of studies [163, 164] is dominated by diboson (or multiboson) production in which leptons are the decay products of the massive weak bosons.

Again, the presence of  $\tau$  leptons is fundamental in some areas of the parameter space since the gauginos may not be as “flavour symmetric” as the corresponding SM bosons. In any case, no significant excess has been observed and the results are used to set limits on the production of gauginos. It should be noted that this kind of final state is sensitive to a different area of the SUSY parameter space, so they are complementary to the search of events in which coloured superpartners are produced and sequentially decay into SM particles.

### 9.1 Gauge-mediated Supersymmetry breaking

After the simplest topologies have been investigated and report negative results regarding the existence of SUSY, other models providing significant differences in the final states need to be considered. A qualitative change is set by models in which SUSY is broken in a hidden sector and communicated via gauge interaction [165], since the LSP is the gravitino and the phenomenology depends on the next-to-lightest SUSY particles (NLSP) because most of the decays go preferably via that particle.

In the cases where such particle is a scalar lepton, usually the scalar  $\tau$ , the final state contains leptons that are easy to identify. Searches by both collaborations [166, 167] show good agreement with expectations in several types of final states.

Other case that is very relevant is when the NLSP is a neutralino, decaying into a gauge boson (usually a photon) and the gravitino. This is also a relatively simple final state, since the presence of photons helps to make the event selection much cleaner. The analysis searching for diphoton and  $E_T^{\text{miss}}$  by CMS [168] observed a good agreement between the observed data and the expected SM backgrounds, as displayed in Fig. 23, where the  $E_T^{\text{miss}}$  distribution in events with two photons is shown, including some possible signals to explicitly shown the sensitivity to a signal in this variable.

Even if the considered final state in models with gauge-mediated SUSY breaking was able to avoid limits set for MSSM-inspired searches, the results are not showing any significant discrepancy that could be attributed to the production of SUSY particles.

### 9.2 Natural SUSY and third generation squarks

After the studies of the more obvious SUSY final states, the obtained limits are moving the SUSY scale to high values so it starts to approach the decoupling with respect to the electroweak scale. Since the motivation for SUSY is to fix problems at this latter scale, new concepts are required to keep the connections between the two scales and, at the same time, avoid the current limits from more inclusive final states.

In this sense the two obvious things is first to keep the neutralino (or equivalent) as the LSP in order to have a Dark Matter candidate that is stable and weakly coupled. Secondly, we need the lightest scalar top to be light enough to keep the divergences in the Higgs mass as smaller as possible. This means  $m(\tilde{t}) \lesssim 400$  GeV. This expression also requires a gluino not far from 2 TeV to avoid a strong correction

on the scalar top mass. With these requirements, all other SUSY particles may have any value, since their influence is much smaller. Therefore current limits on general searches are avoided.

However, this “natural” SUSY becomes only completely natural when other superpartners are associated to the needed ones. For this reason it is not uncommon to have also light scalar bottom quarks or scalar  $\tau$  (as mentioned above). Additionally, the LSP could be a family of degenerated gauginos of several classes. It should be noted that in spite of the reduced number of superparticles involved, the possible final states are very complex due to the involvement of the third generation of fermions.

For example, with the described spectrum, it is feasible to have gluino-pair production as the process with higher cross section. These gluinos decay into quarks and  $E_T^{\text{miss}}$ . In the case the scalar bottom is available, the gluinos may give rise to final states containing  $E_T^{\text{miss}}$  and four bottom quarks, that may be identified as b-jets. This topology is very clean due to the reduced backgrounds and therefore sensitivity may be enhanced by the b-jet requirements, allowing some additional room with respect to the more inclusive limits, where the limitation was the huge backgrounds. The study done by ATLAS [169] shows no hint for anomalous production of multi-b-jets and significant  $E_T^{\text{miss}}$ , a selection sensitive to this final state. Limits in SUSY and other models are set. Regarding the interpretation, it should be noted that this analysis is also sensitive to the decay into top quarks, since also four bottom-quarks appear in the final state.

On the other hand, the case of top and scalar top quarks produced via gluino production is much richer than just the presence of b-jets, due to the large multiplicity of W bosons. It is possible then to identify the events containing four top quarks and significant  $E_T^{\text{miss}}$  in several approaches and with very challenging final states for the SM expectations: analyses in this topic [170, 171] are testing the SM predictions in very specific corners of the phase space, and specifically in regions that were not tested before. Even there the SM predictions provide a very good description of the measurements, which translates into further constraints to SUSY production.

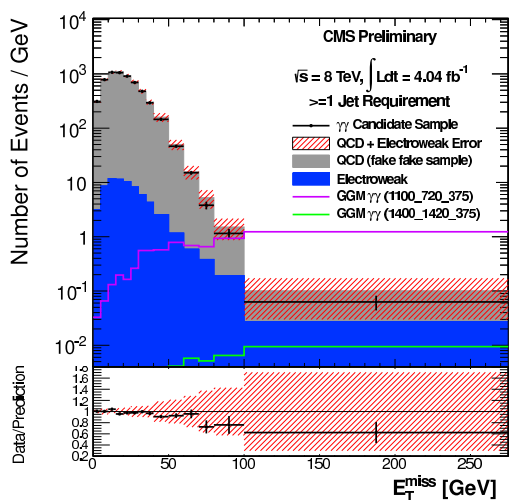
Even if the use of gluino-mediated production allows the use of striking signatures, it is more attractive the direct production of squarks of the third generation which are those strongly motivated to be relatively light, according to “naturalness”. Therefore experiments performed searches of scalar bottom quarks as that in [172] in which the identification of b-jets is fundamental to reduce the SM background. In addition, searches for direct production of scalar top quarks [173–176] still provide enough complexity in the final state to allow several classes of searches. This is seen in summary plots as that displayed in Fig. 24, containing the exclusion areas from several searches of direct production of scalar top quarks.

As the summary plot shows, the several assumptions on the decay and kinematics of the final states allows to exclude large areas of the parameter space. But in summary, the lack of observation of hints for scalar top quarks just bring the scale for SUSY (in this case given by the mass of the scalar top) to higher values, similarly of the results in more inclusive searches. Therefore, it seems that SUSY may not show up in the most obvious way to fix the issues of the SM and particle physics.

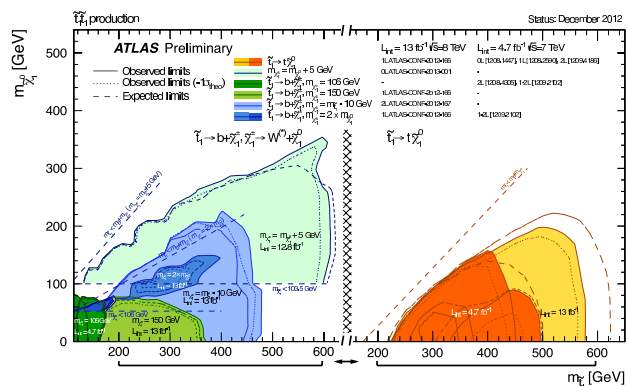
### 9.3 Searches for $R_P$ -Violating SUSY

Although usually it is assumed that  $R_P$  is conserved because it directly provides a Dark Matter candidate, it is obvious that there is no reason a priori why that quantity needs to be conserved. By relaxing the conservation condition it is possible to avoid many of the most stringent limits, since they are usually obtained with the requirement on  $E_T^{\text{miss}}$ , which is inspired by the assumption of conserving  $R_P$ . In addition, the phenomenology becomes much richer due to the possibilities in the spectrum and in the possible interactions. For example with the presence of unusual resonances in the final state (like  $\tilde{\nu}_\tau \rightarrow e\mu$ ).

One general characteristic of the  $R_P$ -Violating signatures is that since all the superpartners decay into SM particles, the final state usually is related with high multiplicity of objects, and involving many



**Fig. 23:** Distribution of the  $E_T^{\text{miss}}$  in diphoton events as measured by the CMS collaboration. Data (dots) are compared to the SM predictions (solid histograms) and to possible models of new physics (lines).



**Fig. 24:** Summary of the limits for scalar tops from the available analyses by ATLAS, drawn in the neutralino (LSP)-stop plane according to the assumptions of all the channels.

different types of them. This also brings the fact that basically every final state is available in  $R_P$ -Violating SUSY due to the rich phenomenology.

As reference analyses, we should mention the multilepton searches, as that by ATLAS [177] looking for anomalous production of events containing 4 or more leptons having either large  $E_T^{\text{miss}}$  or large energy activity (quantified via the concept of effective mass). Reasonable agreement with the small SM expectations has been observed.

Other typical search in the context of  $R_P$ -Violating models is the search for resonances decaying into two leptons of different type, as the one performed by ATLAS in [178], which considers the case of  $e\mu$ ,  $e\tau$  and  $\mu\tau$  resonances. No hint of such states was found and limits were set in the relevant models. It should be noted that the open possibilities in this set of models have the clear disadvantage that the application of the limits is very reduced in comparison with the parameter space.

A last analysis that needs to be discussed is the search for events containing multileptons and identified b-jets, that has been investigated by CMS [179]. The interest of this search is not only about possible presence of new physics, but also since it is sensitive to very rare SM processes, whose observation is as interesting as the search for BSM physics. This includes some of the associated production of top quark and weak bosons mentioned in section 5.2. Although no hint of new physics has been observed, the analysis already probes the sensitivity to the rare SM processes that should be investigated in future datasamples collected at the LHC.

## 10 Future of the LHC experiments and physics

After the running ended in March 2013, the LHC accelerator is currently in a shutdown period which is needed for maintenance and repair work which will allow the running at the highest energy and luminosity conditions. This shutdown will last until 2015 and it is also used by the experiments for additional improvements and work.

The plan after the shutdown is to run for a few years at nominal energy (probably 13 TeV) and collect a sample of  $100 \text{ fb}^{-1}$ . Afterwards a new shutdown is expected to bring the luminosity to the

design value and run for a few more years (2019-2022) to collect additional  $350 \text{ fb}^{-1}$  at a center of mass energy of 14 TeV.

Afterwards a third shutdown will bring the machine to a Phase-2 upgrade that may allow to collect additional  $3000 \text{ fb}^{-1}$  along the next decade. All these data will allow accurate studies for particles and interactions observed during the first runs of the collider. An alternative will be to upgrade the LHC so it may be able to reach higher energies and set a new frontier on the investigated energy scales.

In addition to the improvements by the accelerator, the experiments are getting ready to upgrade their components in order to exploit the possibilities the several stages of the LHC will provide. ATLAS and CMS will need to face new challenges in terms of collection rate, luminosity and radiation and are therefore working on improvements for the DAQ and trigger selection, upgrades of the internal parts of the detectors and replacements of the parts that may be limiting factors in the incoming phases.

In the case of ALICE, the main goal is to have the best possible detector for the run after the second shutdown, in order to get all the reachable information about the heavy ion program of the LHC, hopefully understanding the Quark-Gluon Plasma with unprecedented accuracy and being able to provide enough information for the theoretical characterization of its properties. It is not completely clear yet whether ALICE will be present in the LHC running beyond 2022,

The case of the LHCb is special due to the reduced need for luminosity. The plan is to collect  $5 \text{ fb}^{-1}$  after the current shutdown and then collect  $50 \text{ fb}^{-1}$  during the main part of the main run of the current LHC. As in the case of ALICE, it is not clear whether LHCb will be present in future improvements of the LHC projects, either in terms of luminosity or of new energy regimes.

To summarize, the LHC is planning the future runnings with improved performance in order to provide large amount of data that will yield to important measurements during the several stages of the accelerator. The expected program and the results from the experiments are awaited from the particle-physics community to confirm and improve the results already obtained at the LHC and described in previous sections.

However, it should be remarked that even the current datasample are still providing important and relevant results, as reported on the web pages of the experiments [180].

## 11 Overview and conclusions

The LHC experiments have finished a very successful *Run I* with very important milestones and discoveries in all the topics planned for the program. Confirmations of the SM expectations, measurements of heavy-flavour and top quark physics and results related to heavy-ion collisions have clearly overrule most of the previous achievements due to the new energy frontier, the good performance of the accelerator and the detectors and also to the high quality of the studies.

In the part dedicated to searches for new particles, which is the main goal of the LHC, the current results already made the first big discovery by finding of a new boson having a mass of 125 GeV. For other possible particles expected in extensions of the SM, new limits have been set, highly increasing the constraints for BSM physics.

The properties of the new boson has been measured in the current datasample and they seem to confirm that this boson may be the long-awaited Higgs boson expected in the standard model, the last missing piece of this theory. Further studies are on-going, and others waiting for further running of the LHC, in order to increase the precision of the measurements and confirm this extrem.

Expectations for the future running in 2015 at 13 TeV are getting higher with the increase in reach for possible new particles and also the improved precision of the measurements with the larger data samples expected. Specifically, precision measurements of the properties of the new boson and of other observations that have been accessible at the LHC keep the focus on the LHC results as the more-likely door to the new discoveries in the second half of this decade.



## Acknowledgements

I would like to thank the organizers for giving me the opportunity to take part at the school and for making it possible with their support and help, Furthermore both the organizers and the participants should be recognized for the nice atmosphere and enjoyable time created around the School.

## References

- [1] S. Glashow, Nucl. Phys. **22** (1961) 579.
- [2] S. Weinberg, Phys. Rev. Lett **19** (1967) 1264.
- [3] A. Salam, Nobel Symposium No 8 (Ed. N.Svartholm, Almqvist and Wiksell, Stockholm) (1968) 367.
- [4] F. Englert and R. Brout, Phys. Rev. Lett. **13** (1964) 321;  
P.W. Higgs, Phys. Lett. **12** (1964) 132;  
P.W. Higgs, Phys. Rev. Lett. **13** (1964) 508;  
G.S. Guralnik, C.R. Hagen and T.W.B. Kibble, Phys. Rev. Lett. **13** (1964) 585.
- [5] L.Evans and P. Bryant (eds.), JINST **3** (2008) S08001.
- [6] TOTEM Collaboration, <http://totem.web.cern.ch/Totem/>.
- [7] LHCf Collaboration, <http://home.web.cern.ch/about/experiments/lhcf>.
- [8] MoEDAL Collaboration, <http://moedal.web.cern.ch/>.
- [9] ATLAS Collaboration, JINST **3** (2008) S08003; <http://atlas.web.cern.ch/>.
- [10] CMS Collaboration, JINST **3** (2008) S08004; <http://cms.web.cern.ch/>.
- [11] LHCb Collaboration, JINST **3** (2008) S08005; <http://lhcb.web.cern.ch/>.
- [12] ALICE Collaboration, JINST **3** (2008) S08002; <http://aliceinfo.cern.ch/>.
- [13] ALICE Collaboration, Eur. Phys. J. **C65** (2010) 111.
- [14] ATLAS Collaboration, Phys. Rev. **D83** (2011) 112001.
- [15] CMS Collaboration, JHEP **09** (2010) 091.
- [16] CMS Collaboration, Phys. Lett. **B718** (2013) 795.
- [17] ATLAS Collaboration, Phys. Rev. **D86** (2012) 014022.
- [18] CMS Collaboration, Public Report CMS-PAS-QCD-11-003.
- [19] ATLAS Collaboration, Public Report ATLAS-CONF-2012-128.
- [20] CMS Collaboration, arXiv:1212.6660, submitted to Phys. Rev. **D**.
- [21] ATLAS Collaboration, arXiv:1210.0441, submitted to Eur. Phys. J. **C**.
- [22] CMS Collaboration, JHEP **04** (2012) 084.
- [23] S. Frixione and B.R. Webber, JHEP **06** (2002) 029;  
S. Frixione, P. Nason and B.R. Webber, JHEP **08** (2003) 007.
- [24] ATLAS Collaboration, Public Report ATLAS-CONF-2012-100.
- [25] ATLAS Collaboration, arXiv:1206.5369, submitted to Phys. Rev. **D**.
- [26] CMS Collaboration, JHEP **09** (2012) 029.
- [27] ATLAS Collaboration, Public Report ATLAS-CONF-2011-160.
- [28] LHCb Collaboration, Eur. Phys. J. **C72** (2012) 2168;  
LHCb Collaboration, Eur. Phys. J. **C73** (2013) 2421.
- [29] CMS Collaboration, Public Report CMS-PAS-FSQ-12-010.
- [30] ATLAS Collaboration, arXiv:1211.1913, submitted to JHEP.
- [31] CMS Collaboration, Public Report CMS-PAS-SMP-12-011.
- [32] ATLAS Collaboration, Phys. Rev. **D86** (2012) 072004.

- [33] CMS Collaboration, JHEP **12** (2012) 034.
- [34] LHCb Collaboration, JHEP **06** (2012) 058.
- [35] LHCb Collaboration, JHEP **01** (2013) 111.
- [36] ATLAS Collaboration, Phys. Rev. **D85** (2012) 092002.
- [37] CMS Collaboration, Phys. Lett. **B722** (2013) 238.
- [38] ATLAS Collaboration, Phys. Lett. **B708** (2012) 221.
- [39] CMS Collaboration, Phys. Rev. Lett. **109** (2012) 251801.
- [40] ATLAS Collaboration, Public Report ATLAS-CONF-2012-156.
- [41] CMS Collaboration, Public Report CMS-PAS-EWK-11-013.
- [42] CMS Collaboration, JHEP **06** (2012) 126.
- [43] J. Alwall et al., JHEP **06** (2011) 128.
- [44] CMS Collaboration, Public Report CMS-PAS-SMP-12-003.
- [45] CMS Collaboration, Public Report CMS-PAS-FSQ-12-019.
- [46] CMS Collaboration, Public Report CMS-PAS-EWK-11-009.
- [47] CMS Collaboration, Phys. Lett. **B721** (2013) 190.
- [48] ATLAS Collaboration, Public Report ATLAS-CONF-2012-090.
- [49] ATLAS Collaboration, Public Report ATLAS-CONF-2012-157.
- [50] ATLAS Collaboration, Eur. Phys. J. **C72** (2012) 2173.
- [51] ATLAS Collaboration, Phys. Rev. Lett. **108** (2012) 152001.
- [52] CMS Collaboration, Phys. Rev. Lett. **108** (2012) 252002.
- [53] LHCb Collaboration, Phys. Rev. Lett. **110** (2013) 182001.
- [54] LHCb Collaboration, Phys. Rev. Lett. **110** (2013) 021801.
- [55] LHCb Collaboration, Public Report LHCb-CONF-2012-008.
- [56] LHCb Collaboration, Phys. Lett. **B719** (2013) 318.
- [57] LHCb Collaboration, Phys. Rev. Lett. **110** (2013) 101802.
- [58] N. Cabibbo, Phys. Rev. Lett. **10** (1963) 531;  
M. Kobayashi and T. Maskawa, Prog. Theor. Phys. **49** (1973) 652.
- [59] LHCb Collaboration, Public Report LHCb-CONF-2012-032.
- [60] LHCb Collaboration, Public Report LHCb-CONF-2012-028.
- [61] LHCb Collaboration, Public Report LHCb-CONF-2012-002.
- [62] LHCb Collaboration, Phys. Rev. Lett. **108** (2012) 111602.
- [63] ATLAS Collaboration, Public Report ATLAS-CONF-2012-131.
- [64] CMS Collaboration, Eur. Phys. J. **C73** (2013) 2386.
- [65] ATLAS Collaboration, Public Report ATLAS-CONF-2012-031.
- [66] CMS Collaboration, JHEP **11** (2012) 67.
- [67] ATLAS Collaboration, Public Report ATLAS-CONF-2012-134;  
CMS Collaboration, Public Report CMS-PAS-TOP-12-003.
- [68] ATLAS Collaboration, Public Report ATLAS-CONF-2012-149.
- [69] CMS Collaboration, Public Report CMS-PAS-TOP-12-006;  
CMS Collaboration, Public Report CMS-PAS-TOP-12-007.
- [70] CMS Collaboration, Public Report CMS-PAS-TOP-12-022.
- [71] ATLAS Collaboration, Eur. Phys. J. **C73** (2013) 2261.
- [72] CMS Collaboration, Eur. Phys. J. **C73** (2013) 2339.
- [73] ATLAS Collaboration, Public Report CONF-ATLAS-2012-095;

- CMS Collaboration, Public Report CMS-PAS-TOP-12-001.
- [74] ATLAS Collaboration, Public Report ATLAS-CONF-2011-054.
- [75] CMS Collaboration, Public Report CMS-PAS-TOP-11-008.
- [76] ATLAS Collaboration, Public Report ATLAS-CONF-2011-141.
- [77] CMS Collaboration, Public Report CMS-PAS-TOP-11-031.
- [78] CMS Collaboration, JHEP **06** (2012) 109.
- [79] CMS Collaboration, Public Report CMS-PAS-TOP-12-016.
- [80] ATLAS Collaboration, Public Report ATLAS-CONF-2012-133.
- [81] ATLAS Collaboration, Phys. Rev. Lett. **108** (2012) 212001.
- [82] ATLAS Collaboration, JHEP **06** (2012) 088.
- [83] CMS Collaboration, Public Report CMS-PAS-TOP-11-020.
- [84] CDF Collaboration, T. Aaltonen, et al., Phys. Rev. **D83** (2011) 112003;  
DØ Collaboration, V.M. Abazov, et al., Phys. Rev. **D84** (2011) 112005.
- [85] ATLAS Collaboration, Eur. Phys. J. **C72** (2012) 2039.
- [86] ATLAS Collaboration, Phys. Lett. **B717** (2012) 129.
- [87] ATLAS Collaboration, Public Report ATLAS-CONF-2012-155.
- [88] CMS Collaboration, Public Report CMS-PAS-TOP-12-024.
- [89] CMS Collaboration, Public Report CMS-PAS-TOP-12-011.
- [90] ATLAS Collaboration, Public Report ATLAS-CONF-2012-056.
- [91] ATLAS Collaboration, Public Report ATLAS-CONF-2011-118.
- [92] ATLAS Collaboration, Phys. Lett. **B716** (2012) 142.
- [93] CMS Collaboration, Phys. Rev. Lett. **110** (2013) 022003.
- [94] ALICE Collaboration, Phys. Rev. Lett. **105** (2010) 252302.
- [95] ALICE Collaboration, arXiv:1305.1562.
- [96] CMS Collaboration, Public Report CMS-PAS-HIN-12-011.
- [97] ALICE Collaboration, Phys. Rev. Lett. **110** (2013) 012301.
- [98] ATLAS Collaboration, Phys. Lett. **B719** (2013) 220.
- [99] CMS Collaboration, Phys. Lett. **B718** (2013) 773.
- [100] CMS Collaboration, Public Report CMS-PAS-HIN-12-008.
- [101] ATLAS Collaboration, Phys. Rev. Lett. **110** (2013) 022301.
- [102] CMS Collaboration, Phys. Lett. **B715** (2012) 66.
- [103] ALICE Collaboration, JHEP **09** (2012) 112.
- [104] CMS Collaboration, Phys. Rev. Lett. **109** (2012) 222301.
- [105] CMS Collaboration, Public Report CMS-PAS-HIN-12-014.
- [106] ALICE Collaboration, Phys. Rev. Lett. **109** (2012) 072301.
- [107] CMS collaboration, Phys. Lett. **B718** (2013) 795.
- [108] ALICE Collaboration, Phys. Rev. Lett. **110** (2013) 082302.
- [109] LEP Working Group for Higgs searches, Phys. Lett. **B565** (2003) 61;  
CDF and DØ Collaborations, Public Report FERMILAB-CONF-11-044-E.
- [110] CMS Collaboration, Public Report CMS-PAS-HIG-12-016.
- [111] CMS Collaboration, Public Report CMS-PAS-HIG-12-041.
- [112] CMS Collaboration, Public Report CMS-PAS-HIG-12-042.
- [113] CMS Collaboration, Public Report CMS-PAS-HIG-12-043.
- [114] CMS Collaboration, Public Report CMS-PAS-HIG-12-044.

- [115] ATLAS Collaboration, Public Report ATLAS-CONF-2012-168.
- [116] ATLAS Collaboration, Public Report ATLAS-CONF-2012-169.
- [117] ATLAS Collaboration, Public Report ATLAS-CONF-2012-161.
- [118] ATLAS Collaboration, Public Report ATLAS-CONF-2012-158 .
- [119] ATLAS Collaboration, Public Report ATLAS-CONF-2012-160.
- [120] CMS Collaboration, Public Report CMS-PAS-HIG-12-045.
- [121] CMS Collaboration, PRL 110 (2013) 081803.
- [122] ATLAS Collaboration, Public Report ATLAS-CONF-2012-170.
- [123] ATLAS Collaboration, Public Report ATLAS-CONF-2012-169.
- [124] ATLAS Collaboration, Phys. Lett. **B717** (2012) 70.
- [125] CMS Collaboration, Public Report CMS-PAS-HIG-12-045.
- [126] ATLAS Collaboration, Public Report ATLAS-CONF-2012-094.
- [127] CMS Collaboration, Public Report CMS-PAS-HIG-12-050.
- [128] CMS Collaboration, JHEP **07** (2012) 143.
- [129] ATLAS Collaboration, Public Report ATLAS-CONF-2011-094.
- [130] CMS Collaboration, arXiv:1210.7619, submitted to Phys. Lett. **B**.
- [131] ATLAS Collaboration, Eur. Phys. J. **C72** (2012) 2244.
- [132] ATLAS Collaboration, Public Report ATLAS-CONF-2012-129.
- [133] CMS Collaboration, Public Report CMS-PAS-EXO-12-061.
- [134] CMS Collaboration, Public Report CMS-PAS-EXO-12-059.
- [135] CMS Collaboration, Public Report CMS-PAS-EXO-11-094.
- [136] CMS Collaboration, Public Report CMS-PAS-EXO-12-060.
- [137] ATLAS Collaboration, Phys. Rev. Lett. **109** (2012) 081801.
- [138] ATLAS Collaboration, Phys. Rev. **D85** (2012) 112012.
- [139] CMS Collaboration, Phys. Lett. **B723** (2013) 280.
- [140] ATLAS Collaboration, Public Report ATLAS-CONF-2012-150.
- [141] ATLAS Collaboration, Public Report ATLAS-CONF-2012-146.
- [142] CMS Collaboration, Phys. Lett. **B718** (2012) 329.
- [143] CMS Collaboration, Public Report CMS-PAS-EXO-11-075.
- [144] ATLAS Collaboration, Phys. Lett. **B709** (2012) 158;  
ATLAS Collaboration, Eur. Phys. J. **C72** (2012) 2151.
- [145] CMS Collaboration, Phys. Rev. Lett. **110** (2013) 081801.
- [146] ATLAS Collaboration, Phys. Rev. Lett **110** (2013) 011802.
- [147] ATLAS Collaboration, Public Report ATLAS-CONF-2012-147.
- [148] CMS Collaboration, JHEP **09** (2012) 094.
- [149] CMS Collaboration, Phys. Rrev. Lett. **108** (2012) 111801.
- [150] ATLAS Collaboration, Public Report ATLAS-CONF-2012-130.
- [151] ATLAS Collaboration, Phys. Lett. **B718** (2013) 1284.
- [152] CMS Collaboration, JHEP **05** (2012) 123.
- [153] CMS Collaboration, JHEP **04** (2012) 061.
- [154] CMS Collaboration, Public Report CMS-PAS-EXO-12-026.
- [155] ATLAS Collaboration, JHEP **01** (2013) 131.
- [156] ATLAS Collaboration, JHEP **11** (2012) 138.
- [157] H.E. Haber and G.L. Kane, Phys. Rev. **117** (1985) 75.

- [158] ATLAS Collaboration, Public Report ATLAS-CONF-2012-109.
- [159] CMS Collaboration, Public Report CMS-PAS-SUS-12-028.
- [160] ATLAS Collaboration, Phys. Rev. **D86** (2012) 092002.
- [161] CMS Collaboration, Phys. Rev. **D87** (2013) 072001.
- [162] CMS Collaboration, Public Report CMS-PAS-SUS-11-029.
- [163] ATLAS Collaboration, Public Report ATLAS-CONF-2012-154.
- [164] CMS Collaboration, Public Report CMS-PAS-SUS-12-022.
- [165] G.F. Giudice and R. Rattazzi, Phys. Rept. **322** (1999) 419.
- [166] ATLAS Collaboration, Eur. Phys. J. **C72** (2012) 2215.
- [167] CMS Collaboration, arXiv:1301.3792, submitted to Eur. Phys. J. **C**.
- [168] CMS Collaboration, Public Report CMS-PAS-SUS-12-018.
- [169] ATLAS Collaboration, Public Report ATLAS-CONF-2012-145.
- [170] ATLAS Collaboration, Public Report ATLAS-CONF-2012-151.
- [171] CMS Collaboration, JHEP **03** (2013) 37.
- [172] ATLAS Collaboration, Public Report ATLAS-CONF-2012-165.
- [173] ATLAS Collaboration, Public Report ATLAS-CONF-2013-001.
- [174] ATLAS Collaboration, Eur. Phys. J. **C72** (2012) 2237.
- [175] ATLAS Collaboration, Phys. Rev. Lett. **109** (2012) 211803.
- [176] CMS Collaboration, Public Report CMS-PAS-SUS-11-030.
- [177] ATLAS Collaboration, Public Report ATLAS-CONF-2012-153.
- [178] ATLAS Collaboration, Phys. Lett. **B723** (2013) 15.
- [179] ATLAS Collaboration, Public Report CMS-SUS-12-027.
- [180] ALICE Collaboration, results at <http://aliceinfo.cern.ch/ArtSubmission/publications>;  
ATLAS Collaboration, results at <https://twiki.cern.ch/twiki/bin/view/AtlasPublic>;  
CMS Collaboration, results at <http://cms.web.cern.ch/news/cms-physics-results>;  
LHCb Collaboration, results at <http://lhcbproject.web.cern.ch/lhcbproject/CDS/cgi-bin/index.php>.



## Particle Physics Instrumentation

W. Riegler

CERN, Geneva, Switzerland

### Abstract

This report summarizes a series of three lectures aimed at giving an overview of basic particle detection principles, the interaction of particles with matter, the application of these principles in modern detector systems, as well techniques to read out detector signals in high-rate experiments.

### 1 Introduction

“New directions in science are launched by new tools much more often than by new concepts” is a famous quote from Freeman Dyson’s book *Imagined Worlds*. This is certainly true for the field of particle physics, where new tools such as the cloud chamber, bubble chamber, wire chamber, solid-state detectors, accelerators, etc. have allowed physicists to enter into uncharted territory and to discover unexpected phenomena, the understanding of which has provided a deeper insight into the nature of matter. Looking at all Nobel Prize winners connected to the Standard Model of particle physics, one finds many more experimentalists and “instrumentalists” than theoretically orientated physicists, which is a strong indicator of the essence of new tools for advancing our knowledge.

This report will first discuss a few detector systems in order to illustrate the detector needs and specifications of modern particle physics experiments. Then the interaction of particles with matter, which is of course at the heart of particle detection, will be reviewed. Techniques for tracking with gas detectors and solid-state detectors as well as energy measurement with calorimeters are then elaborated. Finally, the tricks on how to process the signals from these detectors in modern high-rate applications will be discussed.

### 2 Examples of detector systems

The Large Hadron Collider (LHC) experiments ATLAS, CMS, ALICE and LHCb are currently some of the most prominent detectors because of their size, complexity and rate capability. Huge magnet systems, which are used to bend the charged particles in order to measure their momenta, dominate the mechanical structures of these experiments. Proton collision rates of 1 GHz, producing particles and jets of TeV-scale energy, present severe demands in terms of spectrometer and calorimeter size, rate capability and radiation resistance. The fact that only about 100 of the  $10^9$  events per second can be written to disk necessitates highly complex online event selection, i.e. “triggering”. The basic layout of these collider experiments is quite similar. Close to the interaction point there are several layers of pixel detectors that allow the collision vertices to be distinguished and measured with precision on the tens of micrometres level. This also allows short-lived B and D mesons to be identified by their displaced decay vertices. In order to follow the tracks along their curved path up to the calorimeter, a few metres distant from the collision point, one typically uses silicon strip detectors or gas detectors at larger radii. CMS has an “all-silicon tracker” up to the calorimeter, while the other experiments use also gas detectors like so-called straw tubes or a time projection chamber. The trackers are then followed by the electromagnetic and hadron calorimeter, which measures the energy of electrons, photons and hadrons by completely absorbing them in very large amounts of material. The muons, the only particles able to pass through the calorimeters, are then measured at even larger radii by dedicated muon systems. The sequence of vertex detector, tracker for momentum spectrometry, calorimeter for energy measurement followed again by tracking for muons is the classic basic geometry that underlies most collider and even fixed-target experiments. It allows one to distinguish electrons, photons, hadrons and muons and to measure their momenta and energies.

The ALICE and LHCb experiments use a few additional detector systems that allow different hadrons to be distinguished. By measuring the particle's velocity in addition to the momentum, one can identify the mass and therefore the type of hadron. This velocity can be determined by measuring time of flight, the Cerenkov angle or the particle's energy loss. ALICE uses, in addition, the transition radiation effect to separate electrons from hadrons, and has therefore implemented almost all known tricks for particle identification. Another particle detector using all these well-established techniques is the Alpha Magnetic Spectrometer (AMS) that has recently been installed on the *International Space Station*. It is aimed at measuring the primary cosmic-ray composition and energy distribution.

More "exotic" detector geometries are used for neutrino experiments, which demand huge detector masses in order to make the neutrinos interact. The IceCube experiment at the South Pole uses one cubic kilometre of ice as the neutrino detection medium to look for neutrino point sources in the Universe. Neutrinos passing through the Earth from the Northern Hemisphere interact deep down under the ice and the resulting charged particles are travelling upwards at speeds larger than the speed of light in the ice. They therefore produce Cerenkov radiation, which is detected by a series of more than 5000 photon detectors that are immersed into the ice and look downwards. An example of an accelerator-based neutrino experiment is the CERN Neutrino to Gran Sasso (CNGS) beam. A neutrino beam is sent from CERN over a distance of 732 km to the Gran Sasso laboratory in Italy, where some large neutrino detectors are set up. One of them, the OPERA detector, uses more than 150 000 lead bricks as neutrino target. The bricks are built up from alternating sheets of lead and photographic emulsion, which allows tracking with the micrometre precision necessary to identify the tau leptons that are being produced by interaction of tau neutrinos. This "passive" detector is followed by trigger and tracking devices, which detect secondary particles from the neutrino interactions in the lead bricks and identify the bricks where an interesting event has taken place. To analyse the event, the bricks have then to be removed from the assembly and the photographic emulsion must be developed.

These are only a few examples from a large variety of existing detector systems. It is, however, important to bear in mind that there are only a few basic principles of particle interaction with matter that underly all these different detectors. It is therefore worth going through them in detail.

### 3 Basics of particle detection

The Standard Model of particle physics counts 17 particles, namely six quarks, six leptons, photon, gluon, W and Z bosons, and the hypothetical Higgs particle. Quarks, however, are not seen as free particles; rather, they combine into baryons and mesons, of which there are hundreds. How can we therefore distinguish all these different particle types in our detectors? The important fact is that, out of the hundreds of known hadrons, only 27 have a lifetime that is long enough such that they can leave a track  $> 1 \mu\text{m}$  in the detector. All the others decay "on the spot" and can only be identified and reconstructed through kinematic relations of their decay products like the "invariant mass". Out of these 27 particles, 13 have lifetimes that make them decay after a distance between a few hundred micrometres and a few millimetres at GeV energies, so they can be identified by their decay vertices, which are only a short distance from the primary collision vertex (secondary vertex tagging). The 14 remaining particles are the only ones that can actually "fly" though the entire detector, and the following eight are by far the most frequent ones: electron, muon, photon, charged pion, charged kaon, neutral kaon, proton and neutron. The principle task of a particle detector is therefore to identify and measure the energies and momenta of these eight particles.

Their differences in mass, charge and type of interaction are the key to their identification, which will be discussed in detail later. The electron leaves a track in the tracking detector and produces a shower in the electromagnetic (EM) calorimeter. The photon does not leave a track but also produces a shower in the EM calorimeter. The charged pion, charged kaon and the proton show up in the tracker but pass through the EM calorimeter and produce hadron showers in the hadron calorimeter. The neutral kaon and the neutron do not show tracks and shower in the hadron calorimeter. The muon is the only particle than



manages to pass through even the hadron calorimeter and is identified by tracking detectors behind the calorimeters. How to distinguish between pion, kaon and proton is typically the task of specific particle identification (PID) detectors.

#### 4 Interaction of particles with matter

The processes leading to signals in particle detectors are now quite well understood and, as a result of available computing power and simulation programs like GEANT or GARFIELD, one can simulate detector responses to the level of a few percent based on fundamental microphysics processes (atomic and nuclear cross-sections). By knowing the basic principles and performing some “back-of-the-envelope calculations”, it is possible to estimate detector response to the 20–30% level.

It sounds obvious that any device that is to detect a particle must interact with it in some way. In accelerator experiments, however, there is a way to detect neutrinos even if they do not interact in the detector. Since the total momentum of the colliding particles is known, the sum of all momenta of the produced particles must amount to the same number, owing to momentum conservation. If one uses a hermetic detector, the measurement of missing momentum can therefore be used to detect the momentum vector of the neutrino!

The electromagnetic interaction of charged particles with matter lies at the heart of all particle detection. We can distinguish six types of these interactions: atomic excitation, atomic ionization, bremsstrahlung, multiple scattering, Cerenkov radiation and transition radiation. We will discuss them in more detail in the following.

##### 4.1 Ionization and excitation

A charged particle passing through an atom will interact through the Coulomb force with the atomic electrons and the nucleus. The energy transferred to the electrons is about 4000 times larger compared to the energy transferred to the nucleus because of the much higher mass of the nucleus. We can therefore assume that energy is transferred only to the electrons. In a distant encounter between a passing particle and an electron, the energy transfer will be small – the electron will not be liberated from the atom but will just go to an excited state. In a close encounter the energy transfer can be large enough to exceed the binding energy – the atom is ionized and the electron is liberated. The photons resulting from de-excitation of the atoms and the ionization electrons and ions are used in particle detectors to generate signals that can be read out with appropriate readout electronics.

The faster the particle is passing through the material, the less time there is for the Coulomb force to act, and the energy transfer for the non-relativistic regime therefore decreases with particle velocity  $v$  as  $1/v^2$ . If the particle velocity reaches the speed of light, this decrease should stop and stay at a minimum plateau. After a minimum for Lorentz factors  $\gamma = 1/\sqrt{1 - v^2/c^2}$  of  $\approx 3$ , however, the energy loss increases again because the kinematically allowed maximum energy that can be transferred from the incoming particle to the atomic electron is increasing. This rise goes with  $\log \gamma$  and is therefore called the relativistic rise. Bethe and Bloch devised a quantum-mechanical calculation of this energy loss in the 1930s. For ultra-relativistic particles, the very strong transverse field will polarize the material and the energy loss will be slightly reduced.

The energy loss is, in addition, independent of the mass of the incoming particle. Dividing the energy loss by the density of the material, it becomes an almost universal curve for all materials. The energy loss of a particle with  $\gamma \approx 3$  is around  $1-2 \times \rho[\text{g}/\text{cm}^3]$  MeV/cm. Taking iron as an example, the energy for a high-energy particle due to ionization and excitation is about 1 GeV/m. The energy loss is also proportional to the square of the particle charge, so a helium nucleus will deposit four times more energy compared to a proton of the same velocity.

Dividing this energy loss by the ionization energy of the material, we can get a good estimate of the number of electrons and ions that are produced in the material along the track of the passing particle.

Since the energy deposited is a function of the particle's velocity only, we can use it to identify particles: measuring the momentum by the bending in a magnetic field and the velocity from the energy loss, we can determine the mass of the particle in certain momentum regions.

If a particle is stopped in a material, the fact that the energy loss of charged particles increases for smaller velocities results in large energy deposits at the end of the particle track. This is the basis of hadron therapy, where charged particles are used for tumour treatment. These particles deposit a large amount of dose inside the body at the location of the tumour without exposing the overlying tissue to high radiation loads.

This energy loss is, of course, a statistical process, so the actual energy loss will show fluctuations around the average given by the Bethe–Bloch description. This energy-loss distribution was first described by Landau and it shows a quite asymmetric tail towards large values of the energy loss. This large fluctuation of the energy loss is one of the important limiting factors of tracking detector resolution.

## 4.2 Multiple scattering, bremsstrahlung and pair production

The Coulomb interaction of an incoming particle with the atomic nuclei of the detector material results in deflection of the particle, which is called multiple scattering. A particle entering a piece of material perpendicular to the surface will therefore have a probability of exiting at a different angle, which has a Gaussian distribution with a standard deviation that depends on the particle's properties and the material. This standard deviation is inversely proportional to the particle velocity and the particle momentum, so evidently the effect of multiple scattering and related loss of tracking resolution and therefore momentum resolution is worst for low-energy particles. The standard deviation of the angular deflection is, in addition, proportional to the square root of the material thickness, so clearly one wants to use the thinnest possible tracking devices. The material properties are summarized in the so-called radiation length  $X_0$ , and the standard deviation depends on the inverse root of that. Materials with small radiation length are therefore not well suited to the volume of tracking devices. This radiation length  $X_0$  is proportional to  $A/\rho Z^2$  where  $A$ ,  $\rho$  and  $Z$  are the nuclear number, density and atomic number of the material. Tracking systems therefore favour materials with very low atomic number like beryllium for beampipes, carbon fibre and aluminium for support structures, and thin silicon detectors or gas detectors as tracking elements.

The deflection of the charged particle by the nuclei results in acceleration and therefore emission of electromagnetic radiation. This effect is called “bremsstrahlung” and it plays a key role in calorimetric measurements. The energy loss of a particle due to bremsstrahlung is proportional to the particle energy and inversely proportional to the square of the particle mass. Since electrons and positrons are very light, they are the only particles where energy loss due to bremsstrahlung can dominate over energy loss due to ionization at typical present accelerator energies. The energy of a high-energy electron or positron travelling a distance  $x$  in a material decreases as  $\exp(-x/X_0)$ , where  $X_0$  is again the above-mentioned radiation length. The muon, the next lightest particle, has about 200 times the electron mass, so the energy loss from bremsstrahlung is 40 000 times smaller at a given particle energy. A muon must therefore have an energy of more than 400 GeV in order to have an energy loss from bremsstrahlung that dominates over the ionization loss. This fact can be used to distinguish them from other particles, and it is at the basis of electromagnetic calorimetry through a related effect, the so-called pair production.

A high-energy photon has a certain probability of converting into an electron–positron pair in the vicinity of a nucleus. This effect is closely related to bremsstrahlung. The average distance that a high-energy photon travels in a material before converting into an electron–positron pair is also approximately given by the radiation length  $X_0$ . The alternating processes of bremsstrahlung and pair production result in an electromagnetic cascade (shower) of more and more electrons and positrons with increasingly degraded energy until they are stopped in the material by ionization energy loss. We will come back to this in the discussion of calorimetry.

### 4.3 Cerenkov radiation

Charged particles passing through material at velocities larger than the speed of light in the material produce an electromagnetic shock wave that materializes as electromagnetic radiation in the visible and ultraviolet range, the so-called Cerenkov radiation. With  $n$  being the refractive index of the material, the speed of light in the material is  $c/n$ , so the fact that a particle does or does not produce Cerenkov radiation can be used to apply a threshold to its velocity. This radiation is emitted at a characteristic angle with respect to particle direction. This Cerenkov angle  $\Theta_c$  is related to the particle velocity  $v$  by  $\cos \Theta_c = c/nv$ , so by measuring this angle, one can determine the velocity of a charged particle.

### 4.4 Transition radiation

Transition radiation is emitted when a charged particle crosses the boundary between two materials of different permittivity. The probability of emission is proportional to the Lorentz factor  $\gamma$  of the particle and is only appreciable for ultra-relativistic particles, so it is mainly used to distinguish electrons from other hadrons. As an example a particle with  $\gamma = 1000$  has a probability of about 1% to emit a photon on the transition between two materials, so one has to place many layers of material in the form of sheets, foam or fibres in order to produce a measurable amount of radiation. The energy of the emitted photons is in the keV region, so the fact that a charged particle is accompanied by X-rays is used to identify it as an electron or positron.

## 5 Detector principles

In the previous section we have seen how charged particles leave a trail of excited atoms and electron–ion pairs along their track. Now we can discuss how this is used to detect and measure them. We will first discuss detectors based on atomic excitation, so-called scintillators, where the de-excitation produces photons, which are reflected to appropriate photon detectors. Then we discuss gaseous and solid-state detectors based on ionization, where the electrons and ions (holes) drift in electric fields, which induces signals on metallic readout electrodes connected to readout electronics.

### 5.1 Detectors based on scintillation

The light resulting from complex de-excitation processes is typically in the ultraviolet to visible range. The three important classes of scintillators are the noble gases, inorganic crystals and polycyclic hydrocarbons (plastics). The noble gases show scintillation even in their liquid phase. An application of this effect is the liquid argon time projection chamber where the instantaneous light resulting from the passage of the particle can be used to mark the start signal for the drift-time measurement. Inorganic crystals show the largest light yield and are therefore used for precision energy measurement in calorimetry applications and also in nuclear medicine. Plastics constitute the most important class of scintillators owing to their cheap industrial production, robustness and mechanical stability. The light yield of scintillators is typically a few percent of the energy loss. In 1 cm of plastic scintillator, a high-energy particle typically loses 1.5 MeV, of which 15 keV goes into visible light, resulting in about 15 000 photons. In addition to the light yield, the decay time, i.e. the de-excitation time, is an important parameter of the scintillator. Many inorganic crystals such as NaI or CsI show very good light yield, but have decay times of tens, even hundreds, of nanoseconds, so they have to be carefully chosen considering the rate requirements of the experiments. Plastic scintillators, on the other hand, are very fast and have decay times on only the nanosecond scale, and they are therefore often used for precision timing and triggering purposes.

The photons produced inside a scintillator are internally reflected to the sides of the material, where so-called “light guides” are attached to guide the photons to appropriate photon detection devices. A very efficient way to extract the light is to use so-called wavelength shifting fibres, which are attached to the side of the scintillator materials. The light entering the fibre from the scintillator is converted into

a longer wavelength there and it can therefore not reflect back into the scintillator. The light stays in the fibre and is internally reflected to the end, where again the photon detector is placed.

The classic device used to convert these photons into electrical signals is the so-called photomultiplier. A photon hits a photocathode, a material with very small work function, and an electron is liberated. This electron is accelerated in a strong electric field to a dynode, which is made from a material with high secondary electron yield. The one electron hitting the surface will therefore create several electrons, which are again guided to the next dynode, and so on, so that out of the single initial electron one ends up with a sizeable signal of, for example,  $10^7$ – $10^8$  electrons.

In recent years, the use of solid-state photomultipliers, the so-called avalanche photodiodes (APDs), has become very popular, owing to their much lower price and insensitivity to magnetic fields.

## 5.2 Gaseous detectors

A high-energy particle leaves about 80 electron–ion pairs in 1 cm of argon, which is not enough charge to be detected above the readout electronics noise of typically a few hundred to a few thousand electrons, depending on the detector capacitance and electronics design. A sizeable signal is only seen if a few tens or hundreds of particles cross the gas volume at the same time, and in this operational mode such a gas detector, consisting of two parallel metal electrodes with a potential applied to one of them, is called an “ionization chamber”. In order to be sensitive to single particles, a gas detector must have internal electron multiplication. This is accomplished most easily in the wire chamber. Wires of very small diameter, between 10 and 100  $\mu\text{m}$ , are placed between two metallic plates a few millimetres apart. The wires are at a high voltage of a few kilovolts, which results in a very high electric field close to the wire surface. The ionization electrons move towards the thin wires, and, in the strong fields close to the wires, the electrons are accelerated to energies above the ionization energy of the gas, which results in secondary electrons and as a consequence an electron avalanche. Gas gains of  $10^4$ – $10^5$  are typically used, which makes the wire chambers perfectly sensitive to single tracks. In this basic application, the position of the track is therefore given by the position of the wire that carries a signal, so we have a one-dimensional positioning device.

One has to keep in mind that the signal in the wire is not due to the electrons entering wire; rather, the signal is induced while the electrons are moving towards the wire and the ions are moving away from it. Once all charges arrive at the electrode, the signal is terminated. The signals in detectors based on ionization are therefore *induced* on the readout electrodes by the *movement* of the charges. This means that we find signals not only on electrodes that receive charges but also on other electrodes in the detector. For the wire chamber one can therefore segment the metal plates (cathodes) into strips in order to find the second coordinate of the track along the wire direction. In many applications, one does not even read out the wire signals but instead one segments the cathode planes into square or rectangular pads to get the full two-dimensional information from the cathode pad readout. The position resolution is in this case not limited by the pad size. If one uses pad dimensions of the order of the cathode-to-wire distance, one finds signals on a few neighbouring pads, and, by using centre-of-gravity interpolation, one can determine the track position, which is only 1/10 to 1/100 of the pad size. Position resolution down to 50  $\mu\text{m}$  and rate capabilities of hundreds of kHz of particles per  $\text{cm}^2$  per second can be achieved with these devices.

Another way to achieve position resolution that is far smaller than the wire separation is the so-called drift chamber. One determines the time when the particle passes the detector by an external device, which can be a scintillator or the accelerator clock in a collider experiment, and one uses the arrival time of the ionization electrons at the wire as the measure of the distance between the track and the wire. The ATLAS muons system, for instance, uses tubes of 15 mm radius with a central wire, and the measurement of the drift time determines the track position to 80  $\mu\text{m}$  precision.

The choice of the gas for a given gas detector is dominated by the transport properties of electrons

and ions in gases, because these determine the signal and timing characteristics. In order to avoid the ionization electrons getting lost on their way to the readout wires, one can use only gases with very small electronegativity. The main component of detector gases are therefore the noble gases like argon or neon. Other admixtures like hydrocarbons (methane, isobutane) or  $\text{CO}_2$  are also needed in order to “tune” the gas transport properties and to ensure operational stability. Since hydrocarbons were shown to cause severe chamber ageing effects at high rates, the LHC detectors use almost exclusively argon, neon and xenon together with  $\text{CO}_2$  for all wire chambers.

Typical drift velocities of electrons are in the range of 5–10 cm/ $\mu\text{s}$ . The velocity of the ions that are produced in the electron avalanche at the wire and are moving back to the cathodes is about 1000–5000 times smaller than the electron velocity. The movement of these ions produced long signal tails in wire chambers, which have to be properly removed by dedicated filter electronics.

During the past 10–15 years a very large variety of new gas detectors have entered particle physics instrumentation, the so-called micropattern gas detectors like the GEM (gas electron multiplier) or the MICROMEGA (micro mesh gas detector). In these detectors the high fields for electron multiplication are produced by micropattern structures that are realized with photolithographic methods. Their main advantages are rate capabilities far in excess of those achievable in wire chambers, low material budget construction and semi-industrial production possibilities.

### 5.3 Solid-state detectors

In gaseous detectors, a charged particle liberates electrons from the atoms, which are freely bouncing between the gas atoms. An applied electric field makes the electrons and ions move, which induces signals on the metal readout electrodes. For individual gas atoms, the electron energy levels are discrete.

In solids (crystals), the electron energy levels are in “bands”. Inner-shell electrons, in the lower energy bands, are closely bound to the individual atoms and always stay with “their” atoms. However, in a crystal there are energy bands that are still bound states of the crystal, but they belong to the entire crystal. Electrons in these bands and the holes in the lower band can move freely around the crystal, if an electric field is applied. The lowest of these bands is called the “conduction band”.

If the conduction band is filled, the crystal is a conductor. If the conduction band is empty and “far away” from the last filled band, the valence band, the crystal is an insulator. If the conduction band is empty but the distance to the valence band is small, the crystal is called a semiconductor.

The energy gap between the valence band and the conduction band is called the band gap  $E_g$ . The band gaps of diamond, silicon and germanium are 5.5, 1.12 and 0.66 eV, respectively. If an electron in the valence band gains energy by some process, it can be excited into the conduction band and a hole in the valence band is left behind. Such a process can be the passage of a charged particle, but also thermal excitation with a probability proportional to  $\exp(-E_g/kT)$ . The number of electrons in the conduction band therefore increases with temperature, i.e. the conductivity of a semiconductor increases with temperature.

It is possible to treat electrons in the conduction band and holes in the valence band similar to free particles, but with an effective mass different from elementary electrons not embedded in the lattice. This mass is furthermore dependent on other parameters such as the direction of movement with respect to the crystal axis. If we want to use a semiconductor as a detector for charged particles, the number of charge carriers in the conduction band due to thermal excitation must be smaller than the number of charge carriers in the conduction band produced by the passage of a charged particle. Diamond can be used for particle detection at room temperature; silicon and germanium must be cooled, or the free charge carriers must be eliminated by other tricks like “doping”.

The average energy to produce an electron–hole pair for diamond, silicon and germanium, respectively, is 13, 3.6 and 2.9 eV. Compared to gas detectors, the density of a solid is about a factor of 1000 larger than that of a gas, and the energy to produce an electron–hole pair for silicon, for example, is

a factor 7 smaller than the energy to produce an electron–ion pair in argon. The number of primary charges in a silicon detector is therefore about  $10^4$  times larger than in a gas and, as a result, solid-state detectors do not need internal amplification. While, in gaseous detectors, the velocities of electrons and ions differ by a factor of 1000, the velocities of electrons and holes in many semiconductor detectors are quite similar, which results in very short signals of a few tens of nanosecond length.

The diamond detector works like a solid-state ionization chamber. One places diamond of a few hundred micrometres thickness between two metal electrodes and applies an electric field. The very large electron and hole mobilities of diamond result in very fast and short signals, so, in addition to tracking application, the diamond detectors are used as precision timing devices.

Silicon is the most widely used semiconductor material for particle detection. A high-energy particle produces around 33 000 electron–hole pairs in  $300\ \mu\text{m}$  of silicon. At room temperature there are, however,  $1.45 \times 10^{10}$  electron–hole pairs per  $\text{cm}^3$ . To apply silicon as a particle detector at room temperature, one therefore has to use the technique of “doping”. Doping silicon with arsenic makes it an n-type conductor (more electrons than holes); doping silicon with boron makes it a p-type conductor (more holes than electrons). Putting an n-type and p-type conductor in contact realizes a diode.

At a p–n junction the charges are depleted and a zone free of charge carriers is established. By applying a voltage, the depletion zone can be extended to the entire diode, which results in a highly insulating layer. An ionizing particle produces free charge carriers in the diode, which drift in the electric field and therefore induce an electrical signal on the metal electrodes. As silicon is the most commonly used material in the electronics industry, it has one big advantage with respect to other materials, namely highly developed technology.

Strip detectors are a very common application, where the detector is segmented into strips of a few  $50\text{--}150\ \mu\text{m}$  pitch and the signals are read out on the ends by wire bonding the strips to the readout electronics. The other coordinate can then be determined, either by another strip detector with perpendicular orientation, or by implementing perpendicular strips on the same wafer. This technology is widely used at the LHC, and the CMS tracker uses  $445\ \text{m}^2$  of silicon detectors.

In the very-high-multiplicity region close to the collision point, a geometry of crossed strips results in too many “ghost” tracks, and one has to use detectors with a chessboard geometry, so-called pixel detectors, in this region. The major complication is the fact that each of the chessboard pixels must be connected to a separate readout electronics channel. This is achieved by building the readout electronics wafer in the same geometry as the pixel layout and soldering (bump bonding) each of the pixels to its respective amplifier. Pixel systems in excess of 100 million channels are successfully operating at the LHC.

A clear goal of current solid-state detector development is the possibility of integration of the detection element and the readout electronics into a monolithic device.

## 6 Calorimetry

The energy measurement of charged particles by completely absorbing (“stopping”) them is called calorimetry. Electromagnetic (EM) calorimeters measure the energy of electrons and photons. Hadron calorimeters measure the energy of charged and neutral hadrons.

### 6.1 Electromagnetic calorimeters

As discussed above, high-energy electrons suffer significant bremsstrahlung owing to their small mass. The interplay of bremsstrahlung and pair production will develop a single electron or photon into a shower of electrons and positrons. The energy of these shower particles decreases exponentially until all of them are stopped due to ionization loss. The total amount of ionization produced by the electrons and positrons is then a measure of the particle energy. The characteristic length scale of this shower process is called the radiation length  $X_0$ , and in order to fully absorb a photon or electron one typically uses a

thickness of about  $25 X_0$ . One example of such an EM calorimeter at the LHC is the crystal calorimeter of CMS, which uses  $\text{PbW}_4$  crystals. The radiation length  $X_0$  of this crystal is 9 mm, so with a length of 22 cm one can fully absorb the high-energy electron and photon showers. In these crystals the light produced by the shower particles is used as the measure of the energy.

Liquid noble gases are the other prominent materials used for EM calorimetry. In these devices, the total amount of ionization is used as a measure of the energy. The NA48 experiment uses a homogeneous calorimeter of liquid krypton, which has a radiation length of 4.7 cm. Liquid argon has a radiation length of 14 cm, so one would need a depth of 350 cm to fully absorb the EM showers. Since this is not practicable, one interleaves the argon with absorber material of smaller radiation length, such as lead, to allow a more compact design of the calorimeter. Such an alternating assembly of absorber material and active detector material is called a sampling calorimeter. Although the energy resolution of such a device is worse compared to a homogeneous calorimeter, for many applications it is good enough. The ATLAS experiment uses such a liquid argon sampling calorimeter. Other calorimeter types use plastic scintillators interleaved with absorber materials.

The energy resolution of calorimeters improves as  $1/\sqrt{E}$  where  $E$  is the particle energy. This means that the energy measurement becomes “easier” at high-energy colliders. For homogeneous EM calorimeters, energy resolutions of  $\sigma_E/E = 1\%/\sqrt{E \text{ (GeV)}}$  are achieved; typical resolutions of sampling calorimeters are in the range of  $\sigma_E/E = (10\text{--}20\%)/\sqrt{E \text{ (GeV)}}$ .

## 6.2 Hadron calorimeters

While only electrons and photons have small enough masses to produce significant EM bremsstrahlung, there is a similar “strong-interaction bremsstrahlung effect” for hadrons. High-energy hadrons radiate pions in the vicinity of a nucleus, and a cascade of these pions develops, which also fully absorbs the incident hadron, and the total ionization loss of this cascade is used to measure the particle energy. The length scale of this shower development is the so-called hadronic interaction length  $\lambda$ , which is significantly larger than the radiation length  $X_0$ . For iron the radiation length  $X_0$  is 1.7 cm, whereas the hadronic interaction length  $\lambda$  is 17 cm. Hadron calorimeters are therefore significantly larger and heavier than EM calorimeters. The energy resolution of hadron calorimeters is typically worse than that of EM calorimeters because of the more complex shower processes. About 50% of the energy ends up in pions, 20% ends up in nuclear excitation and 30% goes into slow neutrons, which are usually not detected. A fraction of the produced pions consists of  $\pi_0$ , which instantly decay into two photons, which in turn start an EM cascade. The relative fluctuations of all these processes will result in a larger fluctuation of the calorimeter signal and therefore reduced resolution. Hadron calorimeters are also typically realized as sampling calorimeters with lead or steel plates interleaved with scintillators or liquid noble gases. Energy resolutions of  $\sigma_E/E = (50\text{--}100\%)/\sqrt{E \text{ (GeV)}}$  are typical.

## 7 Particle identification

By measuring the trajectory of a particle in a magnetic field, one measures the particle’s momentum, so in order to determine the particle type, i.e. the particle’s mass, one needs an additional measurement. Electrons, positrons and photons can be identified by electromagnetic calorimetry, and muons can be identified by the fact that they traverse large amounts of material without being absorbed. To distinguish between protons, kaons and pions is a slightly more subtle affair, and it is typically achieved by measuring the particle’s velocity in addition to the momentum.

For kinetic energies that are not too far from the rest mass of the particle, the velocity is not yet too close to the speed of light, such that one can measure the velocity by time of flight. With precision timing detectors like scintillators or resistive plate chambers, time resolutions of less than 100 ps are being achieved. For a time-of-flight distance of 1 m, this allows kaon/pion separation up to  $1.5 \text{ GeV}/c$ , and proton/pion separation up to about  $3 \text{ GeV}/c$ .

The energy loss of a particle also measures its velocity, so particle identification up to tens of GeV for pions and protons can be achieved. In gas detectors with pad readout and charge interpolation, the signal pulse height is measured for centre-of-gravity interpolation in view of precision tracking. Since the pulse height is a measure of the energy loss, it can in addition be used for particle identification. Time projection chambers are the best examples of combined tracking and particle identification detectors.

For larger velocities, one can use the measurement of the Cerenkov angle to find the particle velocity. This radiation is emitted at a characteristic angle that is uniquely related to the particle velocity. Using short radiators this angle can be determined simply by measuring the radius of the circle produced by the photons in a plane at a given distance from the radiator. Another technique uses a spherical mirror to project the photons emitted along a longer path onto a plane that also forms a circle. Detectors of this type are called ring imaging Cerenkov detectors (RICH). Since only a “handful” of photons are emitted over typical radiator thicknesses, very efficient photon detectors are the key ingredient to Cerenkov detectors. Using very long gas radiators with very small refractive index, kaon/pion separation up to momenta of 200 GeV/ $c$  has been achieved.

## 8 Signal readout

Many different techniques to make particle tracks visible were developed in the last century. The cloud chamber, the bubble chamber and the photographic emulsion were taking actual pictures of the particle tracks. Nowadays we have highly integrated electronic detectors that allow high particle rates to be processed with high precision. Whereas bubble chambers were almost unbeatable in terms of position resolution (down to a few micrometres) and the ability to investigate very complex decay processes, these detectors were only able to record a few events per second, which is not suitable for modern high-rate experiments. The LHC produces  $10^9$  proton–proton collisions per second, of which, for example, 100 produce W bosons that decay into leptons, 10 produce a top quark pair and 0.1 produce a hypothetical Higgs particle of 100 GeV. Only around 100 of the  $10^9$  events per second can be written to tape, which still results in petabytes of data per year to be analysed. The techniques to reduce the rate from  $10^9$  to 100 Hz by selecting only the “interesting” events is the realm of the so-called trigger and data acquisition. With a bunch crossing time of 25 ns, the particles produced in one collision have not even reached the outer perimeter of the detector when the next collision is already taking place. The synchronization of the data belonging to one single collision is therefore another very challenging task. In order to become familiar with the techniques and vocabulary of trigger and data acquisition, we discuss a few examples.

If, for example, we want to measure temperature, we can use the internal clock of a PC to periodically trigger the measurement. If, on the other hand, we want to measure the energy spectrum of the beta-decay electrons of a radioactive nucleus, we need to use the signal itself to trigger the readout. We can split the detector signal caused by the beta electron and use one path to apply a threshold to the signal, which produces a “logic” pulse that can “trigger” the measurement of the pulse height in the second path. Until this trigger signal is produced, one has to “store” the signal somewhere, which is done in the simplest application by a long cable where the signal can propagate.

If we measure the beta electrons, we cannot distinguish the signals from cosmic particles that are traversing the detector. By building a box around our detector that is made from scintillator, for example, we can determine whether a cosmic particle has entered the detector or whether it was a genuine beta-decay electron. Triggering the readout on the condition of a detector signal in coincidence with the absence of a signal in the scintillator box, we can therefore arrive at a pure beta spectrum sample.

Another example of a simple “trigger” logic is the measurement of the muon lifetime with a stack of three scintillators. Many of the cosmic muons will pass through all three scintillators, but some of them will have lower energy such that they traverse the first one and get stuck in the central one. After a certain time the muon will decay and the decay electron produces a signal in the central and the bottom scintillators. By starting a clock with a signal condition of 1 AND 2 AND NOT 3 and stopping the clock



with NOT 1 AND 2 AND 3, one can measure the lifetime of the muons.

At the LHC experiment some typical trigger signals are high-energy events transverse to the proton beam direction, which signify interesting high-energy parton collisions. High-energy clusters in the calorimeters or high-energy muons are therefore typical trigger signals, which start the detector readout and ship the data to dedicated processing units for further selection refinement.

In order to cope with high rates, one has to find appropriate ways to deal with the “processing” time, i.e. the time while the electronics is busy with reading out the data. This we discuss in the following. First we assume a temperature sensor connected to a PC. The PC has an internal clock, which can be used to periodically trigger the temperature measurement and write the values to disk. The measurement and data storage will take a certain time  $\tau$ , so this “deadtime” limits the maximum acquisition rate. For a deadtime  $\tau = 1$  ms, we have a maximum acquisition rate of  $f = 1/\tau = 1$  kHz.

For the example of the beta spectrum measurement, we are faced with the fact that the events are completely random and it can happen that another beta decay takes place while the acquisition of the previous one is still ongoing. In order to avoid triggering the readout while the acquisition of the previous event is still ongoing, one has to introduce a so-called “busy logic”, which blocks the trigger while the readout is ongoing. Because the time between events typically follows an exponential distribution, there will always be events lost even if the acquisition time is smaller than the average rate of events. In order to collect 99% of the events, one has to overdesign the readout system with a deadtime of only 10% of the average time between events. To avoid this problem, one uses a so-called FIFO (first-in first-out) buffer in the data stream. This buffer receives as input the randomly arriving data and stores them in a queue. The readout of the buffer happens at constant rate, so by properly choosing the depth of the buffer and the readout rate, it is possible to accept all data without loss, even for readout rates close to the average event rate. This transformation from random input to clocked output is called “de-randomization”.

In order to avoid “storing” the signals in long cables, one can also replace them by FIFOs. At colliders, where the bunch crossing comes in regular intervals, the data are stored in so-called front-end pipelines, which sample the signals at the bunch crossing rate and store them until a trigger decision arrives.

The event selection is typically performed at several levels of increasing refinement. The fast trigger decisions in the LHC experiments are performed by specialized hardware on or close to the detector. After a coarse events selection, the rates are typically low enough to allow a more refined selection using dedicated computer farms that do more sophisticated analysis of the events. The increasing computing power, however, drives the concepts of trigger and data acquisition into quite new directions. The concepts for some future high-energy experiments foresee so-called “asynchronous” data-driven readout concepts, where the signal of each detector element receives a time stamp and is then shipped to a computer farm where the event synchronization and events selection is carried out purely by software algorithms.



# Practical Statistics for the LHC

*K. Cranmer*

Center for Cosmology and Particle Physics, Physics Department, New York University, USA

## Abstract

This document is a pedagogical introduction to statistics for particle physics. Emphasis is placed on the terminology, concepts, and methods being used at the Large Hadron Collider. The document addresses both the statistical tests applied to a model of the data and the modeling itself. I expect to release updated versions of this document in the future.

## 1 Introduction

It is often said that the language of science is mathematics. It could well be said that the language of experimental science is statistics. It is through statistical concepts that we quantify the correspondence between theoretical predictions and experimental observations. While the statistical analysis of the data is often treated as a final subsidiary step to an experimental physics result, a more direct approach would be quite the opposite. In fact, thinking through the requirements for a robust statistical statement is an excellent way to organize an analysis strategy.

In these lecture notes<sup>1</sup> I will devote significant attention to the strategies used in high-energy physics for developing a statistical model of the data. This modeling stage is where you inject your understanding of the physics. I like to think of the modeling stage in terms of a conversation. When your colleague asks you over lunch to explain your analysis, you tell a story. It is a story about the signal and the backgrounds – are they estimated using Monte Carlo simulations, a side-band, or some data-driven technique? Is the analysis based on counting events or do you use some discriminating variable, like an invariant mass or perhaps the output of a multivariate discriminant? What are the dominant uncertainties in the rate of signal and background events and how do you estimate them? What are the dominant uncertainties in the shape of the distributions and how do you estimate them? The answer to these questions forms a *scientific narrative*; the more convincing this narrative is the more convincing your analysis strategy is. The statistical model is the mathematical representation of this narrative and you should strive for it to be as faithful a representation as possible.

Once you have constructed a statistical model of the data, the actual statistical procedures should be relatively straight forward. In particular, the statistical tests can be written for a generic statistical model without knowledge of the physics behind the model. The goal of the RooStats project was precisely to provide statistical tools based on an arbitrary statistical model implemented with the RooFit modeling language. While the formalism for the statistical procedures can be somewhat involved, the logical justification for the procedures is based on a number of abstract properties for the statistical procedures. One can follow the logical argument without worrying about the detailed mathematical proofs that the procedures have the required properties. Within the last five years there has been a significant advance in the field's understanding of certain statistical procedures, which has led to some commonalities in the statistical recommendations by the major LHC experiments. I will review some of the most common statistical procedures and their logical justification.

---

<sup>1</sup>These notes borrow significantly from other documents that I am writing contemporaneously; specifically Ref. [1], documentation for HistFactory [2] and the ATLAS Higgs combination.

## 2 Conceptual building blocks for modeling

### 2.1 Probability densities and the likelihood function

This section specifies my notations and conventions, which I have chosen with some care.<sup>2</sup> Our statistical claims will be based on the outcome of an experiment. When discussing frequentist probabilities, one must consider ensembles of experiments, which may either be real, based on computer simulations, or mathematical abstraction.

Figure 1 establishes a hierarchy that is fairly general for the context of high-energy physics. Imagine the search for the Higgs boson, in which the search is composed of several “channels” indexed by  $c$ . Here a channel is defined by its associated event selection criteria, not an underlying physical process. In addition to the number of selected events,  $n_c$ , each channel may make use of some other measured quantity,  $x_c$ , such as the invariant mass of the candidate Higgs boson. The quantities will be called “observables” and will be written in roman letters e.g.  $x_c$ . The notation is chosen to make manifest that the observable  $x$  is frequentist in nature. Replication of the experiment many times will result in different values of  $x$  and this ensemble gives rise to a *probability density function* (pdf) of  $x$ , written  $f(x)$ , which has the important property that it is normalized to unity

$$\int f(x) dx = 1 .$$

In the case of discrete quantities, such as the number of events satisfying some event selection, the integral is replaced by a sum. Often one considers a parametric family of pdfs

$$f(x|\alpha) ,$$

read “ $f$  of  $x$  given  $\alpha$ ” and, henceforth, referred to as a *probability model* or just *model*. The parameters of the model typically represent parameters of a physical theory or an unknown property of the detector’s response. The parameters are not frequentist in nature, thus any probability statement associated with  $\alpha$  is Bayesian.<sup>3</sup> In order to make their lack of frequentist interpretation manifest, model parameters will be written in greek letters, e.g.:  $\mu, \theta, \alpha, \nu$ .<sup>4</sup> From the full set of parameters, one is typically only interested in a few: the *parameters of interest*. The remaining parameters are referred to as *nuisance parameters*, as we must account for them even though we are not interested in them directly.

While  $f(x)$  describes the probability density for the observable  $x$  for a single event, we also need to describe the probability density for a dataset with many events,  $\mathcal{D} = \{x_1, \dots, x_n\}$ . If we consider the events as independently drawn from the same underlying distribution, then clearly the probability density is just a product of densities for each event. However, if we have a prediction that the total number of events expected, call it  $\nu$ , then we should also include the overall Poisson probability for observing  $n$  events given  $\nu$  expected. Thus, we arrive at what statisticians call a marked Poisson model,

$$\mathbf{f}(\mathcal{D}|\nu, \alpha) = \text{Pois}(n|\nu) \prod_{e=1}^n f(x_e|\alpha) , \quad (1)$$

where I use a bold  $\mathbf{f}$  to distinguish it from the individual event probability density  $f(x)$ . In practice, the expectation is often parametrized as well and some parameters simultaneously modify the expected rate and shape, thus we can write  $\nu \rightarrow \nu(\alpha)$ . In `Roofit` both  $f$  and  $\mathbf{f}$  are implemented with a `RooAbsPdf`; where `RooAbsPdf::getVal(x)` always provides the value of  $f(x)$  and depending on `RooAbsPdf::extendMode()` the value of  $\nu$  is accessed via `RooAbsPdf::expectedEvents()`.

<sup>2</sup>As in the case of relativity, notational conventions can make some properties of expressions manifest and help identify mistakes. For example,  $g_{\mu\nu}x^\mu y^\nu$  is manifestly Lorentz invariant and  $x^\mu + y^\nu$  is manifestly wrong.

<sup>3</sup>Note, one can define a conditional distribution  $f(x|y)$  when the joint distribution  $f(x, y)$  is defined in a frequentist sense.

<sup>4</sup>While it is common to write  $s$  and  $b$  for the number of expected signal and background, these are parameters *not* observables, so I will write  $\nu_S$  and  $\nu_B$ . This is one of few notational differences to Ref. [1].

The *likelihood function*  $L(\alpha)$  is numerically equivalent to  $f(x|\alpha)$  with  $x$  fixed – or  $\mathbf{f}(\mathcal{D}|\alpha)$  with  $\mathcal{D}$  fixed. The likelihood function should not be interpreted as a probability density for  $\alpha$ . In particular, the likelihood function does not have the property that it normalizes to unity

$$\int L(\alpha) d\alpha = 1. \quad \text{Not True!}$$

It is common to work with the log-likelihood (or negative log-likelihood) function. In the case of a marked Poisson, we have what is commonly referred to as an extended likelihood [3]

$$-\ln L(\alpha) = \underbrace{\nu(\alpha) - n \ln \nu(\alpha)}_{\text{extended term}} - \sum_{e=1}^n \ln f(x_e) + \underbrace{\ln n!}_{\text{constant}}.$$

To reiterate the terminology, *probability density function* refers to the value of  $f$  as a function of  $x$  given a fixed value of  $\alpha$ ; *likelihood function* refers to the value of  $f$  as a function of  $\alpha$  given a fixed value of  $x$ ; and *model* refers to the full structure of  $f(x|\alpha)$ .

Probability models can be constructed to simultaneously describe several channels, that is several disjoint regions of the data defined by the associated selection criteria. I will use  $e$  as the index over events and  $c$  as the index over channels. Thus, the number of events in the  $c^{\text{th}}$  channel is  $n_c$  and the value of the  $e^{\text{th}}$  event in the  $c^{\text{th}}$  channel is  $x_{ce}$ . In this context, the data is a collection of smaller datasets:  $\mathcal{D}_{\text{sim}} = \{\mathcal{D}_1, \dots, \mathcal{D}_{c_{\text{max}}}\} = \{\{x_{c=1,e=1} \dots x_{c=1,e=n_c}\}, \dots, \{x_{c=c_{\text{max}},e=1} \dots x_{c=c_{\text{max}},e=n_{c_{\text{max}}}}\}\}$ . In RooFit the index  $c$  is referred to as a RooCategory and it is used to inside the dataset to differentiate events associated to different channels or categories. The class RooSimultaneous associates the dataset  $\mathcal{D}_c$  with the corresponding marked Poisson model. The key point here is that there are now multiple Poisson terms. Thus we can write the combined (or simultaneous) model

$$\mathbf{f}_{\text{sim}}(\mathcal{D}_{\text{sim}}|\alpha) = \prod_{c \in \text{channels}} \left[ \text{Pois}(n_c|\nu(\alpha)) \prod_{e=1}^{n_c} f(x_{ce}|\alpha) \right], \quad (2)$$

remembering that the symbol product over channels has implications for the structure of the dataset.

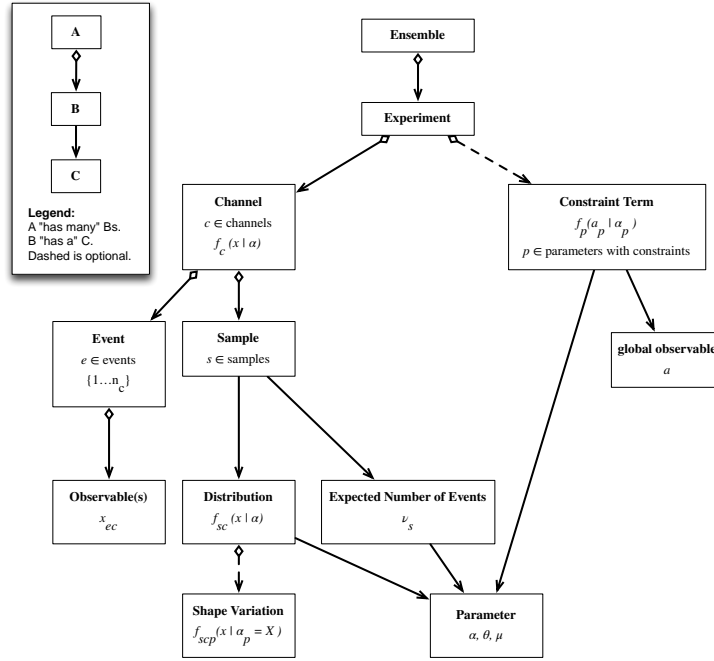
## 2.2 Auxiliary measurements

Auxiliary measurements or control regions can be used to estimate or reduce the effect of systematic uncertainties. The signal region and control region are not fundamentally different. In the language that we are using here, they are just two different channels.

A common example is a simple counting experiment with an uncertain background. In the frequentist way of thinking, the true, unknown background in the signal region is a nuisance parameter, which I will denote  $\nu_B$ .<sup>5</sup> If we call the true, unknown signal rate  $\nu_S$  and the number of events in the signal region  $n_{\text{SR}}$  then we can write the model  $\text{Pois}(n_{\text{SR}}|\nu_S + \nu_B)$ . As long as  $\nu_B$  is a free parameter, there is no ability to make any useful inference about  $\nu_S$ . Often we have some estimate for the background, which may have come from some control sample with  $n_{\text{CR}}$  events. If the control sample has no signal contamination and is populated by the same background processes as the signal region, then we can write  $\text{Pois}(n_{\text{CR}}|\tau\nu_B)$ , where  $n_{\text{CR}}$  is the number of events in the control region and  $\tau$  is a factor used to extrapolate the background from the signal region to the control region. Thus the total probability model can be written  $\mathbf{f}_{\text{sim}}(n_{\text{SR}}, n_{\text{CR}}|\nu_S, \nu_B) = \text{Pois}(n_{\text{SR}}|\nu_S + \nu_B) \cdot \text{Pois}(n_{\text{CR}}|\tau\nu_B)$ . This is a special case of Eq. 2 and is often referred to as the ‘on/off’ problem [4].

Based on the control region alone, one would estimate (or ‘measure’)  $\nu_B = n_{\text{CR}}/\tau$ . Intuitively the estimate comes with an ‘uncertainty’ of  $\sqrt{n_{\text{CR}}}/\tau$ . We will make these points more precise in Sec. 3.1, but

<sup>5</sup>Note, you can think of a counting experiment in the context of Eq. 1 with  $f(x) = 1$ , thus it reduces to just the Poisson term.



**Fig. 1:** A schematic diagram of the logical structure of a typical particle physics probability model and dataset structures.

the important lesson here is that we can use auxiliary measurements (ie.  $n_{CR}$ ) to describe our uncertainty on the nuisance parameter  $\nu_B$  statistically. Furthermore, we have formed a statistical model that can be treated in a frequentist formalism – meaning that if we repeat the experiment many times  $n_{CR}$  will vary and so will the estimate of  $\nu_B$ . It is common to say that auxiliary measurements ‘constrain’ the nuisance parameters. In principle the auxiliary measurements can be every bit as complex as the main signal region, and there is no formal distinction between the various channels.

The use of auxiliary measurements is not restricted to estimating rates as in the case of the on/off problem above. One can also use auxiliary measurements to constrain other parameters of the model. To do so, one must relate the effect of some common parameter  $\alpha_p$  in multiple channels (ie. the signal region and a control regions). This is implicit in Eq. 2.

### 2.3 Frequentist and Bayesian reasoning

The intuitive interpretation of measurement of  $\nu_B$  to be  $n_{CR}/\tau \pm \sqrt{n_{CR}}/\tau$  is that the parameter  $\nu_B$  has a distribution centered around  $n_{CR}/\tau$  with a width of  $\sqrt{n_{CR}}/\tau$ . With some practice you will be able to immediately identify this type of reasoning as Bayesian. It is manifestly Bayesian because we are referring to the probability distribution of a parameter. The frequentist notion of probability of an event is defined as the limit of its relative frequency in a large number of trials. The large number of trials is referred to as an ensemble. In particle physics the ensemble is formed conceptually by repeating the experiment many times. The true values of the parameters, on the other hand, are states of nature, not the outcome of an experiment. The true mass of the  $Z$  boson has no frequentist probability distribution. The existence or non-existence of the Higgs boson has no frequentist probability associated with it. There is a sense in which one can talk about the probability of parameters, which follows from Bayes’s theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{3}$$

Bayes’s theorem is a theorem, so there’s no debating it. It is not the case that Frequentists dispute whether Bayes’s theorem is true. The debate is whether the necessary probabilities exist in the first place. If one can define the joint probability  $P(A, B)$  in a frequentist way, then a Frequentist is perfectly happy using Bayes theorem. Thus, the debate starts at the very definition of probability.

The Bayesian definition of probability clearly can’t be based on relative frequency. Instead, it is based on a degree of belief. Formally, the probability needs to satisfy Kolmogorov’s axioms for probability, which both the frequentist and Bayesian definitions of probability do. One can quantify degree of belief through betting odds, thus Bayesian probabilities can be assigned to hypotheses on states of nature. In practice human’s bets are not generally not ‘coherent’ (see ‘dutch book’), thus this way of quantifying probabilities may not satisfy the Kolmogorov axioms.

Moving past the philosophy and accepting the Bayesian procedure at face value, the practical consequence is that one must supply prior probabilities for various parameter values and/or hypotheses. In particular, to interpret our example measurement of  $n_{CR}$  as implying a probability distribution for  $\nu_B$  we would write

$$\pi(\nu_B|n_{CR}) \propto f(n_{CR}|\nu_B)\eta(\nu_B), \tag{4}$$

where  $\pi(\nu_B|n_{CR})$  is called the *posterior* probability density,  $f(n_{CR}|\nu_B)$  is the likelihood function, and  $\eta(\nu_B)$  is the *prior* probability. Here I have suppressed the somewhat curious term  $P(n_{CR})$ , which can be thought of as a normalization constant and is also referred to as the *evidence*. The main point here is that one can only invert ‘the probability of  $n_{CR}$  given  $\nu_B$ ’ to be ‘the probability of  $\nu_B$  given  $n_{CR}$ ’ if one supplies a prior. Humans are very susceptible to performing this logical inversion accidentally, typically with a uniform prior on  $\nu_B$ . Furthermore, the prior degree of belief cannot be derived in an objective way. There are several formal rules for providing a prior based on formal rules (see Jefferey’s prior and Reference priors), though these are not accurately described as representing a degree of belief. Thus, that style of Bayesian analysis is often referred to as objective Bayesian analysis.

Some useful and amusing quotes on Bayesian and Frequentist reasoning:

*“Using Bayes’s theorem doesn’t make you a Bayesian, **always** using Bayes’s theorem makes you a Bayesian.”* –unknown

*“Bayesians address the questions everyone is interested in by using assumptions that no one believes. Frequentist use impeccable logic to deal with an issue that is of no interest to anyone.”*- Louis Lyons

## 2.4 Consistent Bayesian and Frequentist modeling of constraint terms

Often a detailed probability model for an auxiliary measurement are not included directly into the model. If the model for the auxiliary measurement were available, it could and should be included as an additional channel as described in Sec. 2.2. The more common situation for background and systematic uncertainties only has an estimate, “central value”, or best guess for a parameter  $\alpha_p$  and some notion of uncertainty on this estimate. In this case one typically resorts to including idealized terms into the likelihood function, here referred to as “constraint terms”, as surrogates for a more detailed model of the auxiliary measurement. I will denote this estimate for the parameters as  $a_p$ , to make it manifestly frequentist in nature. In this case there is a single measurement of  $a_p$  per experiment, thus it is referred to as a “global observable” in RooStats. The treatment of constraint terms is somewhat *ad hoc* and discussed in more detail in Sec. 4.1.6. I make it a point to write constraint terms in a manifestly frequentist form  $f(a_p|\alpha_p)$ .

Probabilities on parameters are legitimate constructs in a Bayesian setting, though they will always rely on a prior. In order to distinguish Bayesian pdfs from frequentist ones, greek letters will be used for their distributions. For instance, a generic Bayesian pdf might be written  $\pi(\alpha)$ . In the context of a main

measurement, one might have a prior for  $\alpha_p$  based on some estimate  $a_p$ . In this case, the prior  $\pi(\alpha_p)$  is really a posterior from some previous measurement. It is desirable to write with the help of Bayes theorem

$$\pi(\alpha_p|a_p) \propto L(\alpha_p)\eta(\alpha_p) = f(a_p|\alpha_p)\eta(\alpha_p), \quad (5)$$

where  $\eta(\alpha_p)$  is some more fundamental prior.<sup>6</sup> By taking the time to undo the Bayesian reasoning into an objective pdf or likelihood and a prior we are able to write a model that can be used in a frequentist context. Within RooStats, the care is taken to separately track the frequentist component and the prior; this is achieved with the `ModelConfig` class.

If one can identify what auxiliary measurements were performed to provide the estimate of  $\alpha_p$  and its uncertainty, then it is not a logical fallacy to approximate it with a constraint term, it is simply a convenience. However, not all uncertainties that we deal result from auxiliary measurements. In particular, some theoretical uncertainties are not statistical in nature. For example, uncertainty associated with the choice of renormalization and factorization scales and missing higher-order corrections in a theoretical calculation are not statistical. Uncertainties from parton density functions are a bit of a hybrid as they are derived from data but require theoretical inputs and make various modeling assumptions. In a Bayesian setting there is no problem with including a prior on the parameters associated to theoretical uncertainties. In contrast, in a formal frequentist setting, one should not include constraint terms on theoretical uncertainties that lack a frequentist interpretation. That leads to a very cumbersome presentation of results, since formally the results should be shown as a function of the uncertain parameter. In practice, the groups often read Eq. 5 to arrive at an effective frequentist constraint term.

I will denote the set of parameters with constraint terms as  $\mathbb{S}$  and the global observables  $\mathcal{G} = \{a_p\}$  with  $p \in \mathbb{S}$ . By including the constraint terms explicitly (instead of implicitly as an additional channel) we arrive at the total probability model, which we will not need to generalize any further:

$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G}|\alpha) = \prod_{c \in \text{channels}} \left[ \text{Pois}(n_c|\nu_c(\alpha)) \prod_{e=1}^{n_c} f_c(x_{ce}|\alpha) \right] \cdot \prod_{p \in \mathbb{S}} f_p(a_p|\alpha_p). \quad (6)$$

### 3 Physics questions formulated in statistical language

#### 3.1 Measurement as parameter estimation

One of the most common tasks of the working physicist is to estimate some model parameter. We do it so often, that we often don't realize it. For instance, the sample mean  $\bar{x} = \sum_{e=1}^n x_e/n$  is an estimate for the mean,  $\mu$ , of a Gaussian probability density  $f(x|\mu, \sigma) = \text{Gauss}(x|\mu, \sigma)$ . More generally, an *estimator*  $\hat{\alpha}(\mathcal{D})$  is some function of the data and its value is used to estimate the true value of some parameter  $\alpha$ . There are various abstract properties such as variance, bias, consistency, efficiency, robustness, etc [5]. The bias of an estimator is defined as  $B(\hat{\alpha}) = E[\hat{\alpha}] - \alpha$ , where  $E$  means the expectation value of  $E[\hat{\alpha}] = \int \hat{\alpha}(x)f(x)dx$  or the probability-weighted average. Clearly one would like an unbiased estimator. The variance of an estimator is defined as  $\text{var}[\hat{\alpha}] = E[(\alpha - E[\hat{\alpha}])^2]$ ; and clearly one would like an estimator with the minimum variance. Unfortunately, there is a tradeoff between bias and variance. Physicists tend to be allergic to biased estimators, and within the class of unbiased estimators, there is a well defined minimum variance bound referred to as the Cramér-Rao bound (that is the inverse of the Fisher information, which we will refer to again later).

The most widely used estimator in physics is the maximum likelihood estimator (MLE). It is defined as the value of  $\alpha$  which maximizes the likelihood function  $L(\alpha)$ . Equivalently this value,  $\hat{\alpha}$ , maximizes  $\log L(\alpha)$  and minimizes  $-\log L(\alpha)$ . The most common tool for finding the maximum likelihood estimator is `Minuit`, which conventionally minimizes  $-\log L(\alpha)$  (or any other function) [6]. The jargon is that one 'fits' the function and the maximum likelihood estimate is the 'best fit value'.

<sup>6</sup>Glen Cowan has referred to this more fundamental prior as an 'urprior', which is based on the German use of 'ur' for forming words with the sense of 'proto-, primitive, original'.



When one has a multi-parameter likelihood function  $L(\boldsymbol{\alpha})$ , then the situation is slightly more complicated. The maximum likelihood estimate for the full parameter list,  $\hat{\boldsymbol{\alpha}}$ , is clearly defined. The various components  $\hat{\alpha}_p$  are referred to as the *unconditional maximum likelihood estimates*. In the physics jargon, one says all the parameters are ‘floating’. One can also ask about maximum likelihood estimate of  $\alpha_p$  is with some other parameters  $\alpha_o$  fixed; this is called the *conditional maximum likelihood estimate* and is denoted  $\hat{\alpha}_p(\boldsymbol{\alpha}_o)$ . These are important quantities for defining the profile likelihood ratio, which we will discuss in more detail later. The concept of variance of the estimates is also generalized to the covariance matrix  $cov[\alpha_p, \alpha_{p'}] = E[(\hat{\alpha}_p - \alpha_p)(\hat{\alpha}_{p'} - \alpha_{p'})]$  and is often denoted  $\Sigma_{pp'}$ . Note, the diagonal elements of the covariance matrix are the same as the variance for the individual parameters, ie.  $cov[\alpha_p, \alpha_p] = var[\alpha_p]$ .

In the case of a Poisson model  $Pois(n|\nu)$  the maximum likelihood estimate of  $\nu$  is simply  $\hat{\nu} = n$ . Thus, it follows that the variance of the estimator is  $var[\hat{\nu}] = var[n] = \nu$ . Thus if the true rate is  $\nu$  one expects to find estimates  $\hat{\nu}$  with a characteristic spread around  $\nu$ ; it is in this sense that the measurement has a estimate has some uncertainty or ‘error’ of  $\sqrt{n}$ . We will make this statement of uncertainty more precise when we discuss frequentist confidence intervals.

When the number of events is large, the distribution of maximum likelihood estimates approaches a Gaussian or normal distribution.<sup>7</sup> This does not depend on the pdf  $f(x)$  having a Gaussian form. For small samples this isn’t the case, but this limiting distribution is often referred to as an *asymptotic distribution*. Furthermore, under most circumstances in particle physics, the maximum likelihood estimate approaches the minimum variance or Cramér-Rao bound. In particular, the inverse of the covariance matrix for the estimates is asymptotically given by

$$\Sigma_{pp'}^{-1}(\boldsymbol{\alpha}) = E \left[ - \frac{\partial^2 \log f(x|\boldsymbol{\alpha})}{\partial \alpha_p \partial_{p'}} \Big| \boldsymbol{\alpha} \right], \tag{7}$$

where I have written explicitly that the expectation, and thus the covariance matrix itself, depend on the true value  $\boldsymbol{\alpha}$ . The right side of Eq. 7 is called the (expected) Fisher information matrix. Remember that the expectation involves an integral over the observables. Since that integral is difficult to perform in general, one often uses the observed Fisher information matrix to approximate the variance of the estimator by simply taking the matrix of second derivatives based on the observed data

$$\tilde{\Sigma}_{pp'}^{-1}(\boldsymbol{\alpha}) = - \frac{\partial^2 \log L(\boldsymbol{\alpha})}{\partial \alpha_p \partial_{p'}}. \tag{8}$$

This is what Minuit’s Hesse algorithm<sup>8</sup> calculates to estimate the covariance matrix of the parameters.

### 3.2 Discovery as hypothesis tests

Let us examine the statistical statement associated to the claim of discovery for new physics. Typically, new physics searches are looking for a signal that is additive on top of the background, though in some cases there are interference effects that need to be taken into account and one cannot really talk about ‘signal’ and ‘background’ in any meaningful way. Discovery is formulated in terms of a hypothesis test where the background-only hypothesis plays the role of the null hypothesis and the signal-plus-background hypothesis plays the roll of the alternative. Roughly speaking, the claim of discovery is a statement that the data are incompatible with the background-only hypothesis. Consider the simplest scenario where one is counting events in the signal region,  $n_{SR}$  and expects  $\nu_B$  events from background and  $\nu_S$  events from the putative signal. Then we have the following hypotheses:

<sup>7</sup>There are various conditions that must be met for this to be true, but skip the fine print in these lectures. There are two conditions that are most often violated in particle physics, which will be addressed later.

<sup>8</sup>The matrix is called the Hessian, hence the name.

symbol	statistical name	physics name	probability model
$H_0$	null hypothesis	background-only	$\text{Pois}(n_{SR} \nu_B)$
$H_1$	alternate hypothesis	signal-plus-background	$\text{Pois}(n_{SR} \nu_S + \nu_B)$

In this simple example it's fairly obvious that evidence for a signal shows up as an excess of events and a reasonable way to quantify the compatibility of the observed data  $n_{CR}^0$  and the null hypothesis is to calculate the probability that the background-only would produce at least this many events; the  $p$ -value

$$p = \sum_{n=n_{SR}^0}^{\infty} \text{Pois}(n|\nu_B). \quad (9)$$

If this  $p$ -value is very small, then one might choose to reject the null hypothesis.

Note, the  $p$ -value is *not* to be interpreted as the probability of the null hypothesis given the data – that is a manifestly Bayesian statement. Instead, the  $p$ -value is a statement about the probability to have obtained data with a certain property assuming the null hypothesis.

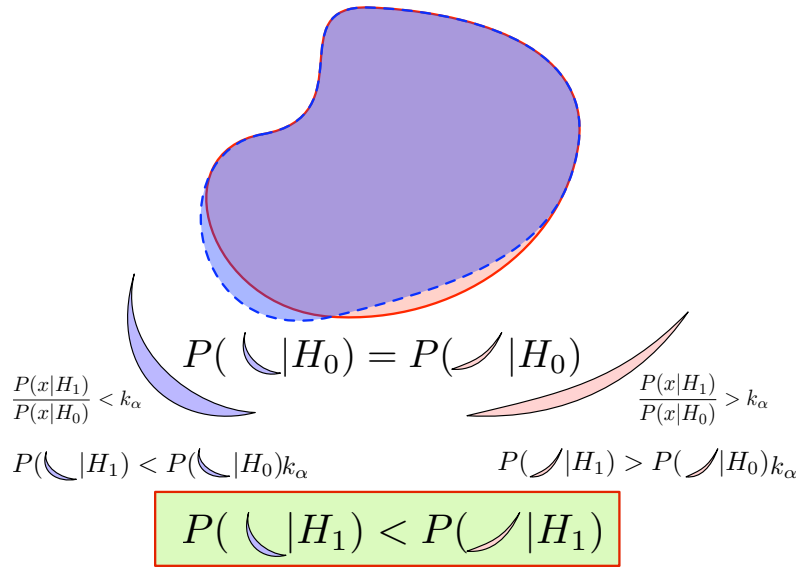
How do we generalize this to more complicated situations? There were really two ingredients in our simple example. The first was the proposal that we would reject the null hypothesis based on the probability for it to produce data at least as extreme as the observed data. The second ingredient was the prescription for what is meant by more discrepant; in this case the possible observations are ordered according to increasing  $n_{SR}$ . One could imagine using difference between observed and expected,  $n_{SR} - \nu_B$ , as the measure of discrepancy. In general, a function that maps the data to a single real number is called a *test statistic*:  $T(\mathcal{D}) \rightarrow \mathbb{R}$ . How does one choose from the infinite number of test statistics?

Neyman and Pearson provided a framework for hypothesis testing that addresses the choice of the test statistic. This setup treats the null and the alternate hypotheses in an asymmetric way. First, one defines an *acceptance region* in terms of a test statistic, such that if  $T(\mathcal{D}) < k_\alpha$  one accepts the null hypothesis. One can think of the  $T(\mathcal{D}) = k_\alpha$  as defining a contour in the space of the data, which is the boundary of this acceptance region. Next, one defines the *size of the test*,  $\alpha$ ,<sup>9</sup> as the probability the null hypothesis will be rejected when it is true (a so-called Type-I error). This is equivalent to the probability under the null hypothesis that the data will not be found in this acceptance region, ie.  $\alpha = P(T(\mathcal{D}) \geq k_\alpha | H_0)$ . Note, it is now clear why there is a subscript on  $k_\alpha$ , since the contour level is related to the size of the test. In contrast, if one accepts the null hypothesis when the alternate is true, it is called a Type-II error. The probability to commit a Type-II error is denoted as  $\beta$  and it is given by  $\beta = P(T(\mathcal{D}) < k_\alpha | H_1)$ . One calls  $1 - \beta$  the *power* of the test. With these definitions in place, one looks for a test statistic that maximizes the power of the test for a fixed test size. This is a problem for the calculus of variations, and sounds like it might be very difficult for complicated probability models.

It turns out that in the case of two simple hypotheses (probability models without any parameters), there is a simple solution! In particular, the test statistic leading to the most powerful test is given by the likelihood ratio  $T_{NP}(\mathcal{D}) = f(\mathcal{D}|H_1)/f(\mathcal{D}|H_0)$ . This result is referred to as the Neyman-Pearson lemma, and I will give an informal proof. We will prove this by considering a small variation to the acceptance region defined by the likelihood ratio. The solid red contour in Fig. 2 represents the rejection region (the complement to the acceptance region) based on the likelihood ratio and the dashed blue contour represents a small perturbation. If we can say that any variation to the likelihood ratio has less power, then we will have proved the Neyman-Pearson lemma. The variation adds (the left, blue wedge) and removes (the right, red wedge) rejection regions. Because the Neyman-Pearson setup requires that both tests have the same size, we know that the probability for the data to be found in the two wedges must be the same under the null hypothesis. Because the two regions are on opposite sides of the contour defined by  $f(\mathcal{D}|H_1)/f(\mathcal{D}|H_0)$ , then we know that the data is less likely to be found in the small region that we added than the small region we subtracted assuming the alternate hypothesis. In other words, there is

<sup>9</sup>Note,  $\alpha$  is the conventional notation for the size of the test, and has nothing to do with a model parameter in Eq. 2.

less probability to reject the null when the alternate is true; thus the test based on the new contour is less powerful.



**Fig. 2:** A graphical proof of the Neyman-Pearson lemma.

How does this generalize for our most general model in Eq. 6 with many free parameters? First one must still define the null and the alternate hypotheses. Typically is done by saying some parameters – the parameters of interest  $\alpha_{\text{poi}}$  – take on specific values takes on a particular value for the signal-plus-background hypothesis and a different value for the background-only hypothesis. For instance, the signal production cross-section might be singled out as the *parameter of interest* and it would take on the value of zero for the background-only and some reference value for the signal-plus-background. The remainder of the parameters are called the *nuisance parameters*  $\alpha_{\text{nuis}}$ . Unfortunately, there is no equivalent to the Neyman-Pearson lemma for models with several free parameters – so called, composite models. Nevertheless, there is a natural generalization based on the profile likelihood ratio.

Remembering that the test statistic  $T$  is a real-valued function of the data, then any particular probability model  $\mathbf{f}_{\text{tot}}(\mathcal{D}|\alpha)$  implies a distribution for the test statistic  $f(T|\alpha)$ . Note, the distribution for the test statistic depends on the value of  $\alpha$ . Below we will discuss how one constructs this distribution, but lets take it as given for the time being. Once one has the distribution, then one can calculate the  $p$ -value is given by

$$p(\alpha) = \int_{T_0}^{\infty} f(T|\alpha)dT = \int \mathbf{f}(\mathcal{D}|\alpha) \theta(T(\mathcal{D}) - T_0) d\mathcal{D} = P(T \geq T_0|\alpha), \quad (10)$$

where  $T_0$  is the value of the test statistic based on the observed data and  $\theta(\cdot)$  is the Heaviside function.<sup>10</sup> Usually the  $p$ -value is just written as  $p$ , but I have written it as  $p(\alpha)$  to make its  $\alpha$ -dependence explicit.

Given that the  $p$ -value depends on  $\alpha$ , how does one decide to accept or reject the null hypothesis? Remembering that  $\alpha_{\text{poi}}$  takes on a specific value for the null hypothesis, we are worried about how the  $p$ -value changes as a function of the nuisance parameters. It is natural to say that one should not reject the null hypothesis if the  $p$ -value is larger than the size of the test *for any value of the nuisance parameters*. Thus, in a frequentist approach one should either present  $p$ -value explicitly as a function of  $\alpha_{\text{nuis}}$  or take

<sup>10</sup>The integral  $\int d\mathcal{D}$  is a bit unusual for a marked Poisson model, because it involves both a sum over the number of events and an integral over the values of  $x_e$  for each of those events.

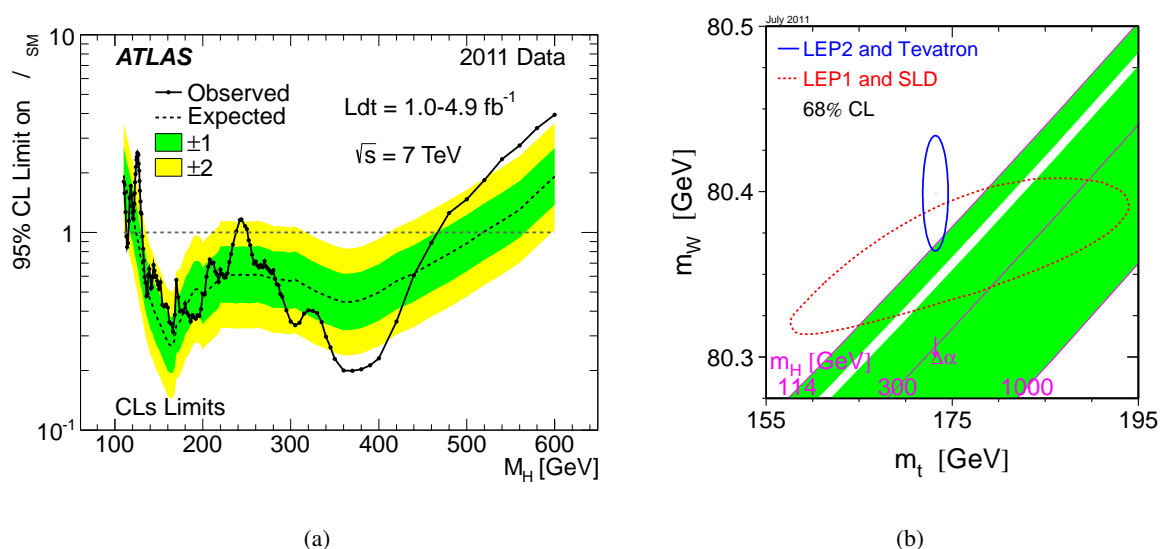
its maximal (or supremum) value

$$p_{\text{sup}}(\alpha_{\text{poi}}) = \sup_{\alpha_{\text{nuis}}} p(\alpha_{\text{nuis}}). \quad (11)$$

As a final note it is worth mentioning that the size of the test, which serves as the threshold for rejecting the null hypothesis, is purely conventional. In most sciences conventional choices of the size are 10%, 5%, or 1%. In particle physics, our conventional threshold for discovery is the infamous  $5\sigma$  criterion – which is a conventional way to refer to  $\alpha = 2.87 \cdot 10^{-7}$ . This is an incredibly small rate of Type-I error, reflecting that claiming the discovery of new physics would be a monumental statement. The origin of the  $5\sigma$  criterion has its roots in the fact that traditionally we lacked the tools to properly incorporate systematics, we fear that there are systematics that may not be fully under control, and we perform many searches for new physics and thus we have many chances to reject the background-only hypothesis. We will return to this in the discussion of the look-elsewhere effect.

### 3.3 Excluded and allowed regions as confidence intervals

Often we consider a new physics model that is parametrized by theoretical parameters. For instance, the mass or coupling of a new particle. In that case we typically want to ask what values of these theoretical parameters are allowed or excluded given available data. Figure 3 shows two examples. Figure 3(a) shows an example with  $\alpha_{\text{poi}} = (\sigma/\sigma_{SM}, M_H)$ , where  $\sigma/\sigma_{SM}$  is the ratio of the production cross-section for the Higgs boson with respect to its prediction in the standard model and  $M_H$  is the unknown Higgs mass parameter in the standard model. All the parameter points above the solid black curve correspond to scenarios for the Higgs boson that are considered ‘excluded at the 95% confidence level’. Figure 3(b) shows an example with  $\alpha_{\text{poi}} = (m_W, m_t)$  where  $m_W$  is the mass of the  $W$ -boson and  $m_t$  is the mass of the top quark. We have discovered the  $W$ -boson and the top quark and measured their masses. The blue ellipse ‘is the 68% confidence level contour’ and all the parameter points inside it are considered ‘consistent with data at the  $1\sigma$  level’. What is the precise meaning of these statements?



**Fig. 3:** Two examples of confidence intervals.

In a frequentist setting, these allowed regions are called *confidence intervals* or *confidence regions*, and the parameter points outside them are considered excluded. Associated with a confidence interval

is a confidence level, i.e. the 95% and 68% confidence level in the two examples. If we repeat the experiments and obtain different data, then these confidence intervals will change. It is useful to think of the confidence intervals as being random in the same way the data are random. The defining property of a 95% confidence interval is that it *covers* the true value 95% of the time.

How can one possibly construct a confidence interval has the desired property, that it *covers* the true value with a specified probability, given that we don't know the true value? The procedure for building confidence intervals is called the Neyman Construction [7], and it is based on 'inverting' a series of hypothesis tests (as described in Sec. 3.2). In particular, for each value of  $\alpha$  in the parameter space one performs a hypothesis test based on some test statistic where the null hypothesis is  $\alpha$ . Note, that in this context, the null hypothesis is changing for each test and generally is not the background-only. If one wants a 95% confidence interval, then one constructs a series of hypothesis test with a size of 5%. The confidence interval  $I(\mathcal{D})$  is constructed by taking the set of parameter points where the null hypothesis is accepted.

$$I(\mathcal{D}) = \{\alpha | P(T(\mathcal{D}) > k_\alpha | \alpha) < \alpha\} , \quad (12)$$

where the final  $\alpha$  and the subscript  $k_\alpha$  refer to the size of the test. Since a hypothesis test with a size of 5% should accept the null hypothesis 95% of the time if it is true, confidence intervals constructed in this way satisfy the defining property. This same property is usually formulated in terms of *coverage*. Coverage is the probability that the interval will contain (cover) the parameter  $\alpha$  when it is true,

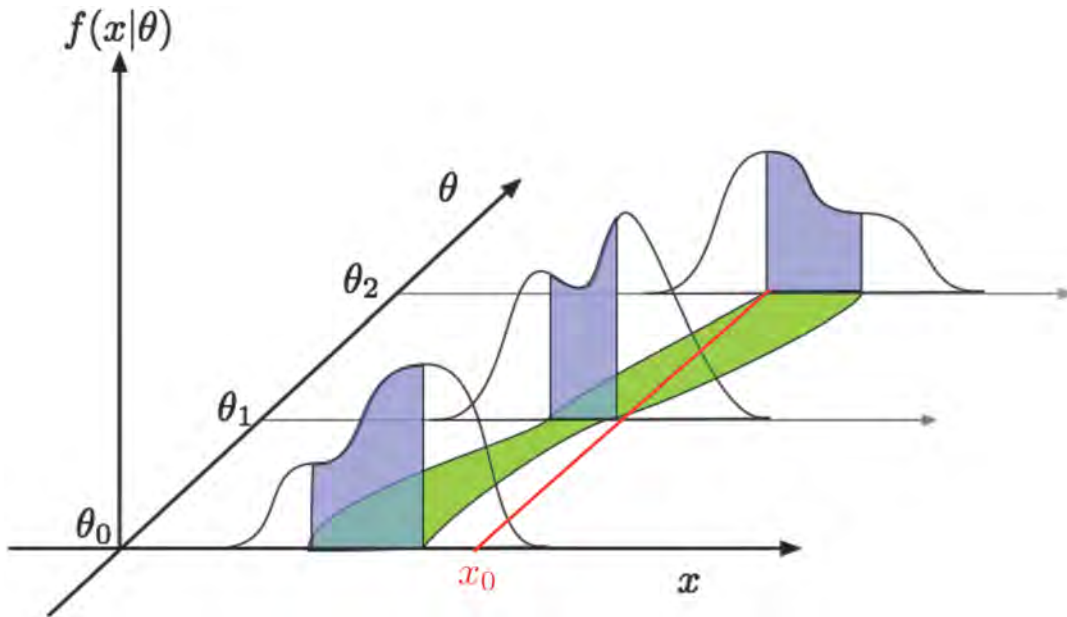
$$\text{coverage}(\alpha) = P(\alpha \in I | \alpha) . \quad (13)$$

The equation above can easily be mis-interpreted as the probability the parameter is in a fixed interval  $I$ ; but one must remember that in evaluating the probability above the data  $\mathcal{D}$ , and, thus, the corresponding intervals produced by the procedure  $I(\mathcal{D})$ , are the random quantities. Note, that coverage is a property that can be quantified for any procedure that produces the confidence intervals  $I$ . Intervals produced using the Neyman Construction procedure are said to "cover by construction"; however, one can consider alternative procedures that may either under-cover or over-cover. Undercoverage means that  $P(\alpha \in I | \alpha)$  is smaller than desired and over-coverage means that  $P(\alpha \in I | \alpha)$  is larger than desired. Note that in general coverage depends on the assumed true value  $\alpha$ .

Since one typically is only interested in forming confidence intervals on the parameters of interest, then one could use the supremum  $p$ -value of Eq. 11. This procedure ensures that the coverage is at least the desired level, though for some values of  $\alpha$  it may over-cover (perhaps significantly). This procedure, which I call the 'full construction', is also computationally very intensive when  $\alpha$  has many parameters as it require performing many hypothesis tests. In the naive approach where each  $\alpha_p$  is scanned in a regular grid, the number of parameter points tested grows exponentially in the number of parameters. There is an alternative approach, which I call the 'profile construction' [8,9] and which statisticians call an 'hybrid resampling technique' [10,11] that is approximate to the full construction, but typically has good coverage properties. We return to the procedures and properties for the different types of Neyman Constructions later.

Figure 4 provides an overview of the classic Neyman construction corresponding to the left panel of Fig. 5. The left panel of Fig. 5 is taken from the Feldman and Cousins's paper [12] where the parameter of the model is denoted  $\mu$  instead of  $\theta$ . For each value of the parameter  $\mu$ , the acceptance region in  $x$  is illustrated as a horizontal bar. Those regions are the ones that satisfy  $T(\mathcal{D}) < k_\alpha$ , and in the case of Feldman-Cousins the test statistic is the one of Eq. 53. This presentation of the confidence belt works well for a simple model in which the data consists of a single measurement  $\mathcal{D} = \{x\}$ . Once one has the confidence belt, then one can immediately find the confidence interval for a particular measurement of  $x$  simply by taking drawing a vertical line for the measured value of  $x$  and finding the intersection with the confidence belt.

Unfortunately, this convenient visualization doesn't generalize to complicated models with many channels or even a single channel marked Poisson model where  $\mathcal{D} = \{x_1, \dots, x_n\}$ . In those more



**Fig. 4:** A schematic visualization of the Neyman Construction. For each value of  $\theta$  one finds a region in  $x$  that satisfies  $\int f(x|\theta)dx$  (blue). Together these regions form a confidence belt (green). The intersection of the observation  $x_0$  (red) with the confidence belt defines the confidence interval  $[\theta_1, \theta_2]$ .

complicated cases, the confidence belt can still be visualized where the observable  $x$  is replaced with  $T$ , the test statistic itself. Thus, the boundary of the belt is given by  $k_\alpha$  vs.  $\mu$  as in the right panel of Fig. 5. The analog to the vertical line in the left panel is now a curve showing how the observed value of the test statistic depends on  $\mu$ . The confidence interval still corresponds to the intersection of the observed test statistic curve and the confidence belt, which clearly satisfies  $T(\mathcal{D}) < k_\alpha$ . For more complicated models with many parameters the confidence belt will have one axis for the test statistic and one axis for each model parameter.

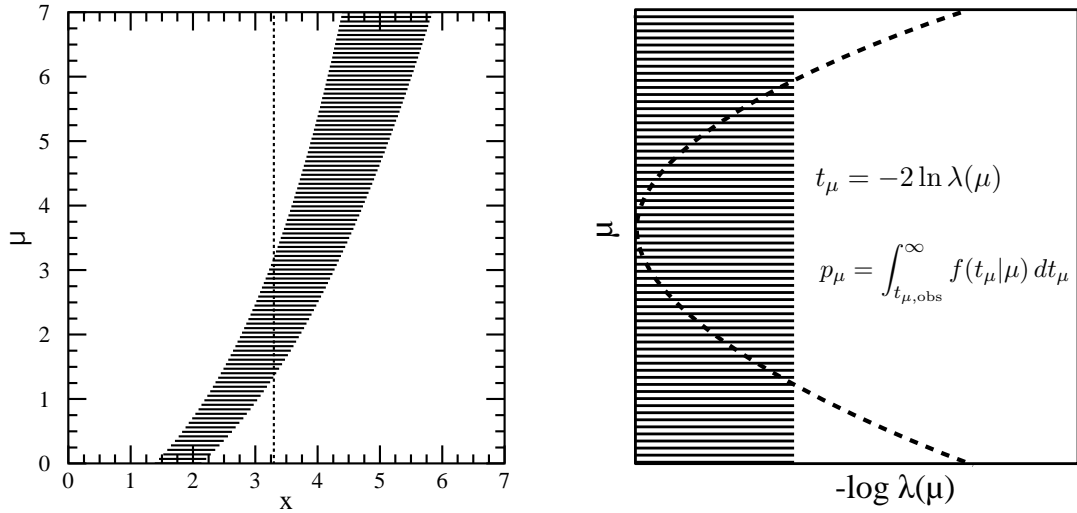
Note, a 95% confidence interval *does not* mean that there is a 95% chance that the true value of the parameter is inside the interval – that is a manifestly Bayesian statement. One can produce a Bayesian *credible interval* with that interpretation; however, that requires a prior probability distribution over the parameters. Similarly, for any fixed interval  $I$  one can compute the Bayesian credibility of the interval

$$P(\alpha \in I|\mathcal{D}) = \frac{\int_I \mathbf{f}(\mathcal{D}|\alpha)\pi(\alpha)d\alpha}{\int \mathbf{f}(\mathcal{D}|\alpha)\pi(\alpha)d\alpha} . \quad (14)$$

#### 4 Modeling and the Scientific Narrative

Now that we have established a general form for a probability model (Eq. 2) and we have translated the basic questions of measurement, discovery, and exclusion into the statistical language we are ready to address the heart of the statistical challenge – building the model. It is difficult to overestimate how important the model building stage is. So many of the questions that are addressed to the statistical experts in the major particle physics collaborations are not really about statistics *per se*, but about model building. In fact, the first question that you are likely to be asked by one of the statistical experts is “what is your model?”

Often people are confused by the question “what is your model?” or simply have not written it down. You simply can’t make much progress on any statistical questions if you haven’t written down a model. Of course, people do usually have some idea for what it is that they want to do. The process of



**Fig. 5:** Two presentations of a confidence belt (see text). Left panel taken from Ref. [12]. Right panel shows a presentation that generalizes to more complicated models.

writing down the model often obviates the answer to the question, reveals some fundamental confusion or assumption in the analysis strategy, or both. As mentioned in the introduction, writing down the model is intimately related with the analysis strategy and it is a good way to organize an analysis effort.

I like to think of the modeling stage in terms of a *scientific narrative*. I find that there are three main narrative elements, though many analyses use a mixture of these elements when building the model. Below I will discuss these narrative elements, how they are translated into a mathematical formulation, and their relative pros and cons.

#### 4.1 Simulation Narrative

The simulation narrative is probably the easiest to explain and produces statistical models with the strongest logical connection to physical theory being tested. We begin with an relation that every particle physicists should know for the rate of events expected from a specific physical process

$$\text{rate} = (\text{flux}) \times (\text{cross section}) \times (\text{efficiency}) \times (\text{acceptance}) , \tag{15}$$

where the cross section is predicted from the theory, the flux is controlled by the accelerator<sup>11</sup>, and the efficiency and acceptance are properties of the detector and event selection criteria. It is worth noting that the equation above is actually a repackaging of a more fundamental relationship. In fact the fundamental quantity that is predicted from first principles in quantum theory is the *scattering probability*  $P(i \rightarrow f) = |\langle i|f \rangle|^2 / (\langle i|i \rangle \langle f|f \rangle)$  inside a box of size  $V$  over some time interval  $T$ , which is then repackaged into the Lorentz invariant form above.

In the simulation narrative the efficiency and acceptance are estimated with computer simulations of the detector. Typically, a large sample of events is generated using Monte Carlo techniques. The Monte Carlo sampling is performed separately for the hard (perturbative) interaction (e.g. MadGraph), the parton shower and hadronization process (e.g. Pythia and Herwig), and the interaction of particles with the detector (e.g. Geant). Note, the efficiency and acceptance depend on the physical process considered, and I will refer to each such process as a *sample* (in reference to the corresponding sample of events generated with Monte Carlo techniques).

<sup>11</sup>In some cases, like cosmic rays, the flux must be estimated since the accelerator is quite far away.

To simplify the notation, I will define the effective cross section,  $\sigma_{\text{eff}}$ , to be the product of the total cross section, efficiency, and acceptance. Thus, the total number of events expected to be selected for a given scattering process,  $\nu$ , is the product of the time-integrated flux or time-integrated luminosity,  $\lambda$ , and the effective cross section

$$\nu = \lambda \sigma_{\text{eff}}. \quad (16)$$

I use  $\lambda$  here instead of the more common  $L$  to avoid confusion with the likelihood function and because when we incorporate uncertainty on the time-integrated luminosity it will be a parameter of the model for which I have chosen to use greek letters.

If we did not need to worry about detector effects and we could measure the final state perfectly, then the distribution for any observable  $x$  would be given by

$$\text{(idealized)} \quad f(x) = \frac{1}{\sigma_{\text{eff}}} \frac{d\sigma_{\text{eff}}}{dx}. \quad (17)$$

Of course, we do need to worry about detector effects and we incorporate them with the detector simulation discussed above. From the Monte Carlo sample of events<sup>12</sup>  $\{x_1, \dots, x_N\}$  we can estimate the underlying distribution  $f(x)$  simply by creating a histogram. If we want we can write the histogram based on  $B$  bins centered at  $x_b$  with bin width  $w_b$  explicitly as

$$\text{(histogram)} \quad f(x) \approx h(x) = \sum_{i=1}^N \sum_{b=1}^B \frac{\theta(|x_i - x_b|/w_b)}{N} \frac{\theta(|x - x_b|/w_b)}{w_b}, \quad (18)$$

where the first Heaviside function accumulates simulated events in the bin and the second selects the bin containing the value of  $x$  in question. Histograms are the most common way to estimate a probability density function based on a finite sample, but there are other possibilities. The downsides of histograms as an estimate for the distribution  $f(x)$  is that they are discontinuous and have dependence on the location of the bin boundaries. A particularly nice alternative is called kernel estimation [13]. In this approach, one places a kernel of probability  $K(x)$  centered around each event in the sample:

$$\text{(kernel estimate)} \quad f(x) \approx \hat{f}_0(x) = \frac{1}{N} \sum_{i=1}^N K\left(\frac{x - x_i}{h}\right). \quad (19)$$

The most common choice of the kernel is a Gaussian distribution, and there are results for the optimal width of the kernel  $h$ . Equation 19 is referred to as the fixed kernel estimate since  $h$  is common for all the events in the sample. A second order estimate or adaptive kernel estimation provides better performance when the distribution is multimodal or has both narrow and wide features [13].

#### 4.1.1 The multi-sample mixture model

So far we have only considered a single interaction process, or sample. How do we form a model when there are several scattering processes contributing to the total rate and distribution of  $x$ ? From first principles of quantum mechanics we must add these different processes together. Since there is no physical meaning to label individual processes that interfere quantum mechanically, I will consider all such processes as a single sample. Thus the remaining set of samples that do not interfere simply add incoherently. The total rate is simply the sum of the individual rates

$$\nu_{\text{tot}} = \sum_{s \in \text{samples}} \nu_s \quad (20)$$

---

<sup>12</sup>Here I only consider unweighted Monte Carlo samples, but the discussion below can be generalized for weighted Monte Carlo samples.



and the total distribution is a weighted sum called a *mixture model*

$$f(x) = \frac{1}{\nu_{\text{tot}}} \sum_{s \in \text{samples}} \nu_s f_s(x), \quad (21)$$

where the subscript  $s$  has been added to the equations above for each such sample. With these two ingredients we can construct our marked Poisson model of Eq. 1 for a single channel, and we can simply repeat this for several disjoint event selection requirements to form a multi-channel simultaneous model like Eq. 2. In the multi-channel case we will give the additional subscript  $c \in \text{channels}$  to  $\nu_{cs}$ ,  $f_{cs}(x)$ ,  $\nu_{c,\text{tot}}$ , and  $f_c(x)$ . However, at this point, our model has no free parameters  $\alpha$ .

#### 4.1.2 Incorporating physics parameters into the model

Now we want to parametrize our model interns of some physical parameters  $\alpha$ , such as those that appear in the Lagrangian of a some theory. Changing the parameters in the Lagrangian of a theory will in general change both the total rate  $\nu$  and the shape of the distributions  $f(x)$ . In principle, we can repeat the procedure above for each value of these parameters  $\alpha$  to form  $\nu_{cs}(\alpha)$  and  $f_{cs}(x|\alpha)$  for each sample and selection channel, and, thus, from  $\mathbf{f}_{\text{sim}}(\mathcal{D}|\alpha)$ . In practice, we need to resort to some interpolation strategy over the individual parameter points  $\{\alpha_i\}$  where we have Monte Carlo samples. We will return to these interpolation strategies later.

In some case the only effect of the parameter is to scale the rate of some scattering process  $\nu_s(\alpha)$  without changing its distribution  $f_s(x|\alpha)$ . Furthermore, the scaling is often known analytically, for instance, a coupling constants produce a linear relationship like  $\nu(\alpha_p) = \xi\alpha_p + \nu_0$ . In such cases, interpolation is not necessary and the parametrization of the likelihood function is straightforward.

Note, not all physics parameters need be considered parameters of interest. There may be a free physics parameter that is not directly of interest, and as such it would be considered a nuisance parameter.

##### 4.1.2.1 An example, the search for the standard model Higgs boson

In the case of searches for the standard model Higgs boson, the only free parameter in the Lagrangian is  $m_H$ . Once  $m_H$  is specified the rates and the shapes for each of the scattering processes (combinations of production and decay modes) are specified by the theory. Of course, as the Higgs boson mass changes the distributions do change so we do need to worry about interpolating the shapes  $f(x|m_H)$ . However the results are often presented as a *raster scan* over  $m_H$ , where one fixes  $m_H$  and then asks about the rate of signal events from the Higgs boson scattering process. With  $m_H$  fixed this is really a simple hypothesis test between background-only and signal-plus-background<sup>13</sup>, but we usually choose to construct a parametrized model that does not directly correspond to any theory. In this case the parameter of interest is some scaling of the rate with respect to the standard model prediction,  $\mu = \sigma/\sigma_{\text{SM}}$ , such that  $\mu = 0$  is the background-only situation and  $\mu = 1$  is the standard model prediction. Furthermore, we usually use this global  $\mu$  factor for each of the production and decay modes even though essentially all theories of physics beyond the standard model would modify the rates of the various scattering processes differently. Figure 3 shows confidence intervals on  $\mu$  for fixed values of  $m_H$ . Values below the solid black curve are not excluded (since an arbitrarily small signal rate cannot be differentiated from the background-only and this is a one-sided confidence interval).

#### 4.1.3 Incorporating systematic effects

The parton shower, hadronization, and detector simulation components of the simulation narrative are based on phenomenological models that have many adjustable parameters. These parameters are nui-

<sup>13</sup>Note that  $H \rightarrow WW$  interferes with “background-only”  $WW$  scattering process. For low Higgs boson masses, the narrow Higgs width means this interference is negligible. However, at high masses the interference effect is significant and we should really treat these two processes together as a single sample.

sance parameters included in our master list of parameters  $\alpha$ . The changes in the rates  $\nu(\alpha)$  and shapes  $f(x|\alpha)$  due to these parameters lead to systematic uncertainties<sup>14</sup>. We have already eluded to how one can deal with the presence of nuisance parameters in hypothesis testing and confidence intervals, but here we are focusing on the modeling stage. In principle, we deal with modeling of these nuisance parameters in the same way as the physics parameters, which is to generate Monte Carlo samples for several choices of the parameters  $\{\alpha_i\}$  and then use some interpolation strategy to form a continuous parametrization for  $\nu(\alpha)$ ,  $f(x|\alpha)$ , and  $\mathbf{f}_{\text{sim}}(\mathcal{D}|\alpha)$ . In practice, there are many nuisance parameters associated to the parton shower, hadronization, and detector simulation so this becomes a multi-dimensional interpolation problem<sup>15</sup>. This is one of the most severe challenges for the simulation narrative.

Typically, we don't map out the correlated effect of changing multiple  $\alpha_p$  simultaneously. Instead, we have some nominal settings for these parameters  $\alpha^0$  and then vary each individual parameter 'up' and 'down' by some reasonable amount  $\alpha_p^\pm$ . So if we have  $N_P$  parameters we typically have  $1 + 2N_P$  variations of the Monte Carlo sample from which we try to form  $\mathbf{f}_{\text{sim}}(\mathcal{D}|\alpha)$ . This is clearly not an ideal situation and it is not hard to imagine cases where the combined effect on the rate and shapes cannot be factorized in terms of changes from the individual parameters.

What is meant by "vary each individual parameter 'up' and 'down' by some reasonable amount" in the paragraph above? The nominal choice of the parameters  $\alpha^0$  is usually based on experience, test beam studies, Monte Carlo 'tunings', etc.. These studies correspond to auxiliary measurements in the language used in Sec. 2.2 and Sec. 2.4. Similarly, these parameters typically have some maximum likelihood estimates and standard uncertainties from the auxiliary measurements as described in Sec. 3.1. Thus our complete model  $\mathbf{f}_{\text{tot}}(\mathcal{D}|\alpha)$  of Eq. 6 should not only deal with parametrizing the effect of changing each  $\alpha_p$  but also include either a constraint term  $f_p(a_p|\alpha_p)$  or an additional channel that describes a more complete probability model for the auxiliary measurement.

Below we will consider a specific interpolation strategy and a few of the most popular conventions for constraint terms. However, before moving on it is worth emphasizing that while, naively, the matrix element associated to a perturbative scattering amplitude has no free parameters (beyond the physics parameters discussed above), fixed order perturbative calculations do have residual scale dependence. This type of *theoretical uncertainty* has no auxiliary measurement associated with it even in principle, thus it really has no frequentist description. This was discussed briefly in Sec. 2.4. In contrast, the parton density functions are the results of auxiliary measurements and the groups producing the parton density function sets spend time providing sensible multivariate constraint terms for those parameters. However, those measurements also have uncertainties due to parametrization choices and theoretical uncertainties, which are not statistical in nature. In short we must take care in ascribing constraint terms to theoretical uncertainties and measurements that have theoretical uncertainties<sup>16</sup>.

#### 4.1.4 Tabulating the effect of varying sources of uncertainty

The treatment of systematic uncertainties is subtle, particularly when one wishes to take into account the correlated effect of multiple sources of systematic uncertainty across many signal and background samples. The most important conceptual issue is that we separate the source of the uncertainty (for instance the uncertainty in the calorimeter's response to jets) from its effect on an individual signal or background sample (eg. the change in the acceptance and shape of a  $W$ +jets background). In particular, the same source of uncertainty has a different effect on the various signal and background samples. The effect of these 'up' and 'down' variations about the nominal predictions  $\nu_s(\alpha^0)$  and  $f_{sb}(x|\alpha^0)$  is quantified by dedicated studies. The result of these studies can be arranged in tables like those below. The main purpose of the HistFactory XML schema is to represent these tables. And HistFactory is a tool that can convert these tables into our master model  $\mathbf{f}_{\text{tot}}(\mathcal{D}|\alpha)$  of Eq. 6 implemented as a RooAbsPdf

<sup>14</sup>Systematic uncertainty is arguably a better term than systematic *error*.

<sup>15</sup>This is sometimes referred to as 'template morphing'

<sup>16</sup>"Note that I deliberately called them theory *errors*, not uncertainties." – Tilman Plehn

with a `ModelConfig` to make it compatible with `RooStats` tools. The convention used by `HistFactory` is related to our notation via

$$\nu_s(\boldsymbol{\alpha})f_s(x|\boldsymbol{\alpha}) = \eta_s(\boldsymbol{\alpha})\sigma_s(x|\boldsymbol{\alpha}) \quad (22)$$

where  $\eta_s(\boldsymbol{\alpha})$  represents relative changes in the overall rate  $\nu(\boldsymbol{\alpha})$  and  $\sigma_s(x|\boldsymbol{\alpha})$  includes both changes to the rate and the shape  $f(x|\boldsymbol{\alpha})$ . This choice is one of convenience because histograms are often not normalized to unity, but instead in code rate information. As the name implies, `HistFactory` works with histograms, so instead of writing  $\sigma_s(x|\boldsymbol{\alpha})$  the table is written as  $\sigma_{sb}(\boldsymbol{\alpha})$ , where  $b$  is a bin index. To compress the notation further,  $\eta_{p=1,s=1}^+$  and  $\sigma_{psb}^\pm$  represent the value of when  $\alpha_p = \alpha_p^\pm$  and all other parameters are fixed to their nominal values. Thus we arrive at the following tabular form for models built on the simulation narrative based on histograms with individual nuisance parameters varied one at a time:

Syst	Sample 1	...	Sample N
Nominal Value	$\eta_{s=1}^0 = 1$	...	$\eta_{s=N}^0 = 1$
$p=OverallSys\ 1$	$\eta_{p=1,s=1}^+, \eta_{p=1,s=1}^-$	...	$\eta_{p=1,s=N}^+, \eta_{p=1,s=N}^-$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$p=OverallSys\ M$	$\eta_{p=M,s=1}^+, \eta_{p=M,s=1}^-$	...	$\eta_{p=M,s=N}^+, \eta_{p=M,s=N}^-$
Net Effect	$\eta_{s=1}(\boldsymbol{\alpha})$	...	$\eta_{s=N}(\boldsymbol{\alpha})$

**Table 1:** Tabular representation of sources of uncertainties that produce a correlated effect in the normalization individual samples (eg. `OverallSys`). The  $\eta_{ps}^+$  represent histogram when  $\alpha_s = 1$  and are inserted into the `High` attribute of the `OverallSys` XML element. Similarly, the  $\eta_{ps}^-$  represent histogram when  $\alpha_s = -1$  and are inserted into the `Low` attribute of the `OverallSys` XML element. Note, this does not imply that  $\eta^+ > \eta^-$ , the  $\pm$  superscript correspond to the variation in the source of the systematic, not the resulting effect.

Syst	Sample 1	...	Sample N
Nominal Value	$\sigma_{s=1,b}^0$	...	$\sigma_{s=N,b}^0$
$p=HistoSys\ 1$	$\sigma_{p=1,s=1,b}^+, \sigma_{p=1,s=1,b}^-$	...	$\sigma_{p=1,s=N,b}^+, \sigma_{p=1,s=N,b}^-$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$p=HistoSys\ M$	$\sigma_{p=M,s=1,b}^+, \sigma_{p=M,s=1,b}^-$	...	$\sigma_{p=M,s=N,b}^+, \sigma_{p=M,s=N,b}^-$
Net Effect	$\sigma_{s=1,b}(\boldsymbol{\alpha})$	...	$\sigma_{s=N,b}(\boldsymbol{\alpha})$

**Table 2:** Tabular representation of sources of uncertainties that produce a correlated effect in the normalization and shape individual samples (eg. `HistoSys`). The  $\sigma_{psb}^+$  represent histogram when  $\alpha_s = 1$  and are inserted into the `HighHist` attribute of the `HistoSys` XML element. Similarly, the  $\sigma_{psb}^-$  represent histogram when  $\alpha_s = -1$  and are inserted into the `LowHist` attribute of the `HistoSys` XML element.

#### 4.1.5 Interpolation Conventions

For each sample, one can interpolate and extrapolate from the nominal prediction  $\eta_s^0 = 1$  and the variations  $\eta_{ps}^\pm$  to produce a parametrized  $\eta_s(\boldsymbol{\alpha})$ . Similarly, one can interpolate and extrapolate from the nominal shape  $\sigma_{sb}^0$  and the variations  $\sigma_{psb}^\pm$  to produce a parametrized  $\sigma_{sb}(\boldsymbol{\alpha})$ . We choose to parametrize  $\alpha_p$  such that  $\alpha_p = 0$  is the nominal value of this parameter,  $\alpha_p = \pm 1$  are the “ $\pm 1\sigma$  variations”. Needless to say, there is a significant amount of ambiguity in these interpolation and extrapolation procedures and they must be handled with care. Bellow are some of the interpolation strategies supported by `HistFactory`. These are all ‘vertical’ style interpolation treated independently per-bin. Four interpolation strategies are described below and can be compared in Fig 6. The interested reader is invited to look at alternative ‘horizontal’ interpolation strategies, such as the one developed by Alex Read in Ref. [14]

(the RooFit implementation is called `RooIntegralMorph`) and Max Baak's `RoomomentMorph`. These horizontal interpolation strategies are better suited for features moving, such as the location of an invariant mass bump changing with the hypothesized mass of a new particle..

### Piecewise Linear (InterpCode=0)

The piecewise-linear interpolation strategy is defined as

$$\eta_s(\boldsymbol{\alpha}) = 1 + \sum_{p \in \text{Syst}} I_{\text{lin.}}(\alpha_p; 1, \eta_{sp}^+, \eta_{sp}^-) \quad (23)$$

and for shape interpolation it is

$$\sigma_{sb}(\boldsymbol{\alpha}) = \sigma_{sb}^0 + \sum_{p \in \text{Syst}} I_{\text{lin.}}(\alpha_p; \sigma_{sb}^0, \sigma_{psb}^+, \sigma_{psb}^-) \quad (24)$$

with

$$I_{\text{lin.}}(\alpha; I^0, I^+, I^-) = \begin{cases} \alpha(I^+ - I^0) & \alpha \geq 0 \\ \alpha(I^0 - I^-) & \alpha < 0 \end{cases} \quad (25)$$

PROS: This approach is the most straightforward of the interpolation strategies.

CONS: It has two negative features. First, there is a kink (discontinuous first derivative) at  $\alpha = 0$  (see Fig 6(b-d)), which can cause some difficulties for numerical minimization packages such as `Minuit`. Second, the interpolation factor can extrapolate to negative values. For instance, if  $\eta^- = 0.5$  then we have  $\eta(\alpha) < 0$  when  $\alpha < -2$  (see Fig 6(c)).

Note that one could have considered the simultaneous variation of  $\alpha_p$  and  $\alpha_{p'}$  in a multiplicative way. The multiplicative accumulation is not an option currently.

Note that this is the default convention for  $\sigma_{sb}(\boldsymbol{\alpha})$  (ie. `HistoSys`).

### Piecewise Exponential (InterpCode=1)

The piecewise exponential interpolation strategy is defined as

$$\eta_s(\boldsymbol{\alpha}) = \prod_{p \in \text{Syst}} I_{\text{exp.}}(\alpha_p; 1, \eta_{sp}^+, \eta_{sp}^-) \quad (26)$$

and for shape interpolation it is

$$\sigma_{sb}(\boldsymbol{\alpha}) = \sigma_{sb}^0 \prod_{p \in \text{Syst}} I_{\text{exp.}}(\alpha_p; \sigma_{sb}^0, \sigma_{psb}^+, \sigma_{psb}^-) \quad (27)$$

with

$$I_{\text{exp.}}(\alpha; I^0, I^+, I^-) = \begin{cases} (I^+/I^0)^\alpha & \alpha \geq 0 \\ (I^-/I^0)^{-\alpha} & \alpha < 0 \end{cases} \quad (28)$$

PROS: This approach ensures that  $\eta(\alpha) \geq 0$  (see Fig 6(c)) and for small response to the uncertainties it has the same linear behavior near  $\alpha \sim 0$  as the piecewise linear interpolation (see Fig 6(a)).

CONS: It has two negative features. First, there is a kink (discontinuous first derivative) at  $\alpha = 0$ , which can cause some difficulties for numerical minimization packages such as `Minuit`. Second, for large uncertainties it develops a different linear behavior compared to the piecewise linear interpolation. In particular, even if the systematic has a symmetric response (ie.  $\eta^+ - 1 = 1 - \eta^-$ ) the interpolated response will develop a kink for large response to the uncertainties (see Fig 6(c)).

Note that the one could have considered the simultaneous variation of  $\alpha_p$  and  $\alpha_{p'}$  in an additive way, but this is not an option currently.

Note, that when paired with a Gaussian constraint on  $\alpha$  this is equivalent to linear interpolation and a log-normal constraint in  $\ln(\alpha)$ . This is the default strategy for normalization uncertainties  $\eta_s(\alpha)$  (ie. OverallSys ) and is the standard convention for normalization uncertainties in the LHC Higgs Combination Group. In the future, the default may change to the Polynomial Interpolation and Exponential Extrapolation described below.

**Polynomial Interpolation and Exponential Extrapolation (InterpCode=4)**

The strategy of this interpolation option is to use the piecewise exponential extrapolation as above with a polynomial interpolation that matches  $\eta(\alpha = \pm\alpha_0)$ ,  $d\eta/d\alpha|_{\alpha=\pm\alpha_0}$ , and  $d^2\eta/d\alpha^2|_{\alpha=\pm\alpha_0}$  and the boundary  $\pm\alpha_0$  is defined by the user (with default  $\alpha_0 = 1$ ).

$$\eta_s(\alpha) = \prod_{p \in \text{Syst}} I_{\text{poly|exp.}}(\alpha_p; 1, \eta_{sp}^+, \eta_{sp}^-, \alpha_0) \tag{29}$$

with

$$I_{\text{poly|exp.}}(\alpha; I^0, I^+, I^-, \alpha_0) = \begin{cases} (I^+/I_0)^\alpha & \alpha \geq \alpha_0 \\ 1 + \sum_{i=1}^6 a_i \alpha^i & |\alpha| < \alpha_0 \\ (I^-/I_0)^{-\alpha} & \alpha \leq -\alpha_0 \end{cases} \tag{30}$$

and the  $a_i$  are fixed by the boundary conditions described above.

PROS: This approach avoids the kink (discontinuous first and second derivatives) at  $\alpha = 0$  (see Fig 6(b-d)), which can cause some difficulties for numerical minimization packages such as Minuit. This approach ensures that  $\eta(\alpha) \geq 0$  (see Fig 6(c)).

**Note:** This option is not available in ROOT 5.32.00, but is available for normalization uncertainties (OverallSys) in the subsequent patch releases. In future releases, this may become the default.

**4.1.6 Consistent Bayesian and Frequentist modeling**

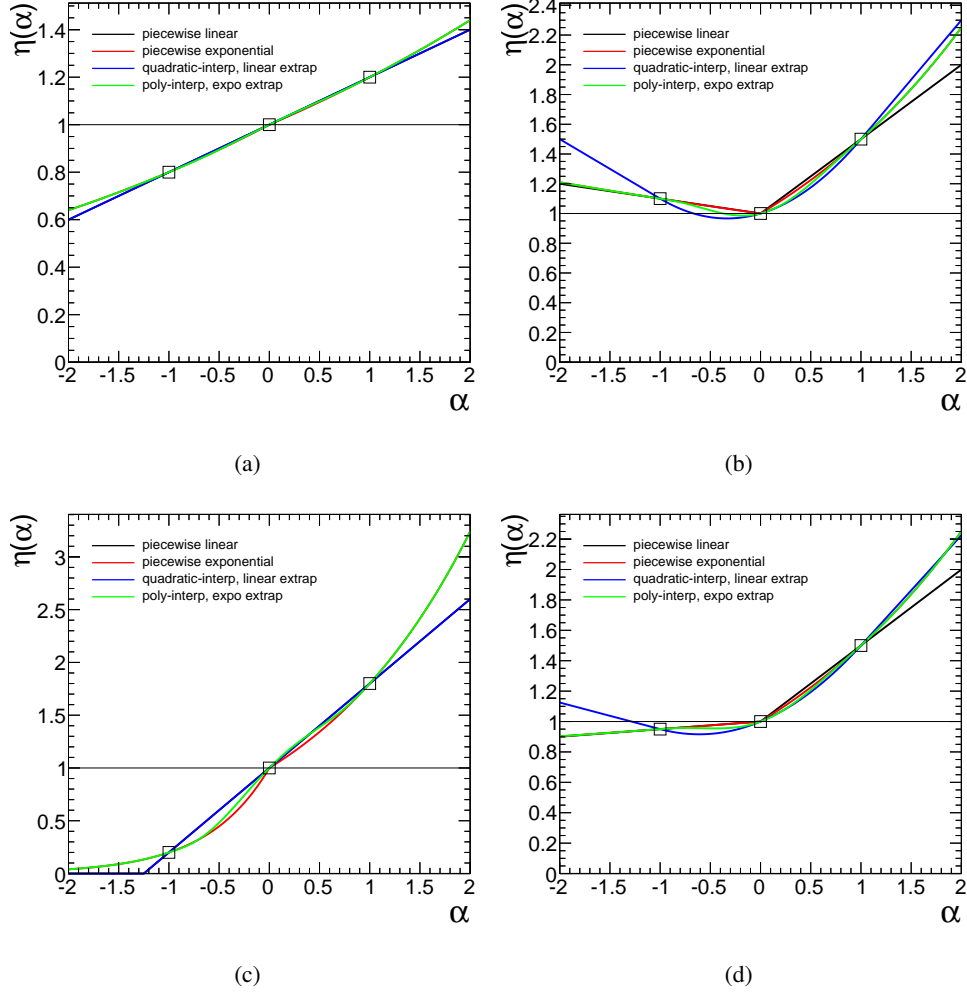
The variational estimates  $\eta^\pm$  and  $\sigma^\pm$  typically correspond to so called “ $\pm 1\sigma$  variations” in the source of the uncertainty. Here we are focusing on the source of the uncertainty, not its affect on rates and shapes. For instance, we might say that the jet energy scale has a 10% uncertainty.<sup>17</sup> This is common jargon, but what does it mean? The most common interpretation of this statement is that the uncertain parameter  $\alpha_p$  (eg. the jet energy scale) has a Gaussian distribution. However, this way of thinking is manifestly Bayesian. If the parameter was estimated from an auxiliary measurement, then it is the PDF for that measurement that we wish to include into our probability model. In the frequentist way of thinking, the jet energy scale has an unknown true value and upon repeating the experiment many times the auxiliary measurements estimating the jet energy scale would fluctuate randomly about this true value. To aid in this subtle distinction, we use greek letters for the parameters (eg.  $\alpha_p$ ) and roman letters for the auxiliary measurements  $a_p$ . Furthermore, we interpret the “ $\pm 1\sigma$ ” variation in the frequentist sense, which leads to the constraint term  $f_p(a_p|\alpha_p)$ . Then, we can pair the resulting likelihood with some prior on  $\alpha_p$  to form a Bayesian posterior if we wish according to Eq. 5.

It is often advocated that a “log-normal” or “gamma” distribution for  $\alpha_p$  is more appropriate than a gaussian constraint [15]. This is particularly clear in the case of bounded parameters and large uncertainties. Here we must take some care to build a probability model that can maintain a consistent interpretation in Bayesian a frequentist settings. Table 3 summarizes a few consistent treatments of the frequentist pdf, the likelihood function, a prior, and the resulting posterior.

Finally, it is worth mentioning that the uncertainty on some parameters is not the result of an auxiliary measurement – so the constraint term idealization, it is not just a convenience, but a real conceptual

---

<sup>17</sup>Without loss of generality, we choose to parametrize  $\alpha_p$  such that  $\alpha_p = 0$  is the nominal value of this parameter,  $\alpha_p = \pm 1$  are the “ $\pm 1\sigma$  variations”.



**Fig. 6:** Comparison of the three interpolation options for different  $\eta^\pm$ . (a)  $\eta^- = 0.8$ ,  $\eta^+ = 1.2$ , (b)  $\eta^- = 1.1$ ,  $\eta^+ = 1.5$ , (c)  $\eta^- = 0.2$ ,  $\eta^+ = 1.8$ , and (d)  $\eta^- = 0.95$ ,  $\eta^+ = 1.5$

PDF	Likelihood $\propto$	Prior $\pi_0$	Posterior $\pi$
$G(a_p \alpha_p, \sigma_p)$	$G(\alpha_p a_p, \sigma_p)$	$\pi_0(\alpha_p) \propto \text{const}$	$G(\alpha_p a_p, \sigma_p)$
$\text{Pois}(n_p \tau_p, \beta_p)$	$P_\Gamma(\beta_p A = \tau_p; B = 1 + n_p)$	$\pi_0(\beta_p) \propto \text{const}$	$P_\Gamma(\beta_p A = \tau_p; B = 1 + n_p)$
$P_{\text{LN}}(n_p \beta_p, \sigma_p)$	$\beta_p \cdot P_{\text{LN}}(\beta_p n_p, \sigma_p)$	$\pi_0(\beta_p) \propto \text{const}$	$P_{\text{LN}}(\beta_p n_p, \sigma_p)$
$P_{\text{LN}}(n_p \beta_p, \sigma_p)$	$\beta_p \cdot P_{\text{LN}}(\beta_p n_p, \sigma_p)$	$\pi_0(\beta_p) \propto 1/\beta_p$	$P_{\text{LN}}(\beta_p n_p, \sigma_p)$

**Table 3:** Table relating consistent treatments of PDF, likelihood, prior, and posterior for nuisance parameter constraint terms.

leap. This is particularly true for theoretical uncertainties from higher-order corrections or renormalization and factorization scale dependence. In these cases a formal frequentist analysis would not include a constraint term for these parameters, and the result would simply depend on their assumed values. As this is not the norm, we can think of reading Table 3 from right-to-left with a subjective Bayesian prior  $\pi(\alpha)$  being interpreted as coming from a fictional auxiliary measurement.

#### 4.1.6.1 Gaussian Constraint

The Gaussian constraint for  $\alpha_p$  corresponds to the familiar situation. It is a good approximation of the auxiliary measurement when the likelihood function for  $\alpha_p$  from that auxiliary measurement has a Gaussian shape. More formally, it is valid when the maximum likelihood estimate of  $\alpha_p$  (eg. the best fit value of  $\alpha_p$ ) has a Gaussian distribution. Here we can identify the maximum likelihood estimate of  $\alpha_p$  with the global observable  $a_p$ , remembering that it is a number that is extracted from the data and thus its distribution has a frequentist interpretation.

$$G(a_p|\alpha_p, \sigma_p) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left[-\frac{(a_p - \alpha_p)^2}{2\sigma_p^2}\right] \quad (31)$$

with  $\sigma_p = 1$  by default. Note that the PDF of  $a_p$  and the likelihood for  $\alpha_p$  are positive for all values.

#### 4.1.6.2 Poisson (“Gamma”) constraint

When the auxiliary measurement is actually based on counting events in a control region (eg. a Poisson process), a more accurate to describe the auxiliary measurement with a Poisson distribution. It has been shown that the truncated Gaussian constraint can lead to undercoverage (overly optimistic) results, which makes this issue practically relevant [4]. Table 3 shows that a Poisson PDF together with a uniform prior leads to a gamma posterior, thus this type of constraint is often called a “gamma” constraint. This is a bit unfortunate since the gamma distribution is manifestly Bayesian and with a different choice of prior, one might not arrive at a gamma posterior. When dealing with the Poisson constraint, it is no longer convenient to work with our conventional scaling for  $\alpha_p$  which can be negative. Instead, it is more natural to think of the number of events measured in the auxiliary measurement  $n_p$  and the mean of the Poisson parameter. This information is not usually available, instead one usually has some notion of the relative uncertainty in the parameter  $\sigma_p^{\text{rel}}$  (eg. a the jet energy scale is known to 10%). In order to give some uniformity to the different uncertainties of this type and think of relative uncertainty, the nominal rate is factored out into a constant  $\tau_p$  and the mean of the Poisson is given by  $\tau_p\alpha_p$ .

$$\text{Pois}(n_p|\tau_p\alpha_p) = \frac{(\tau_p\alpha_p)^{n_p} e^{-\tau_p\alpha_p}}{n_p!} \quad (32)$$

Here we can use the fact that  $\text{Var}[n_p] = \sqrt{\tau_p\alpha_p}$  and reverse engineer the nominal auxiliary measurement

$$n_p^0 = \tau_p = (1/\sigma_p^{\text{rel}})^2. \quad (33)$$

where the superscript 0 is to remind us that  $n_p$  will fluctuate in repeated experiments but  $n_p^0$  is the value of our measured estimate of the parameter.

One important thing to keep in mind is that there is only one constraint term per nuisance parameter, so there must be only one  $\sigma_p^{\text{rel}}$  per nuisance parameter. This  $\sigma_p^{\text{rel}}$  is related to the fundamental uncertainty in the source and we cannot infer this from the various response terms  $\eta_{ps}^{\pm}$  or  $\sigma_{pub}^{\pm}$ .

Another technical difficulty is that the Poisson distribution is discrete. So if one were to say the relative uncertainty was 30%, then we would find  $n_p^0 = 11.11\dots$ , which is not an integer. Rounding  $n_p$  to the nearest integer while maintaining  $\tau_p = (1/\sigma_p^{\text{rel}})^2$  will bias the maximum likelihood estimate of  $\alpha_p$  away from 1. To avoid this, one can use the gamma distribution, which generalizes more continuously with

$$P_{\Gamma}(\alpha_p|A = \tau_p, B = n_p - 1) = A(A\alpha_p)^B e^{-A\alpha_p} / \Gamma(B). \quad (34)$$

This approach works fine for likelihood fits, Bayesian calculations, and frequentist techniques based on asymptotic approximations, but it does not offer a consistent treatment of the pdf for the global observable  $n_p$  that is needed for techniques based on Monte Carlo sampling.

#### 4.1.6.3 Log-normal constraint

From Eadie et al., “The log-normal distribution represents a random variable whose logarithm follows a normal distribution. It provides a model for the error of a process involving many small multiplicative errors (from the Central Limit Theorem). It is also appropriate when the value of an observed variable is a random proportion of the previous observation.” [15, 16]. This logic of multiplicative errors applies to the the measured value, not the parameter. Thus, it is natural to say that there is some auxiliary measurement (global observable) with a log-normal distribution. As in the gamma/Poisson case above, let us again say that the global observable is  $n_p$  with a nominal value

$$n_p^0 = \tau_p = (1/\sigma_p^{\text{rel}})^2. \quad (35)$$

Then the conventional choice for the corresponding log-normal distribution is

$$P_{\text{LN}}(n_p|\alpha_p, \kappa_p) = \frac{1}{\sqrt{2\pi} \ln \kappa} \frac{1}{n_p} \exp \left[ -\frac{\ln(n_p/\alpha_p)^2}{2(\ln \kappa_p)^2} \right] \quad (36)$$

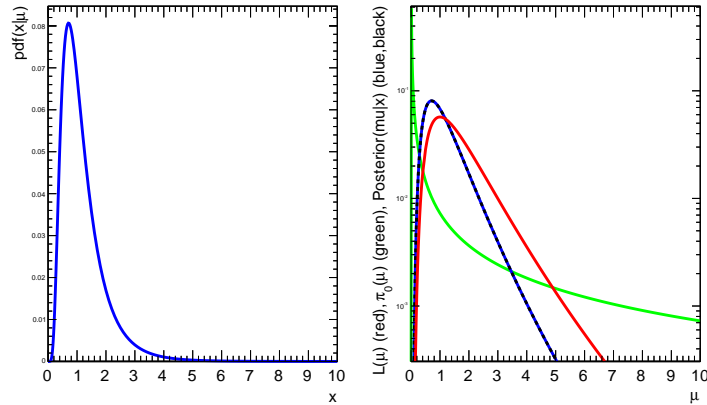
while the likelihood function is (blue curve in Fig. 7(a)).

$$L(\alpha_p) = \frac{1}{\sqrt{2\pi} \ln \kappa} \frac{1}{n_p} \exp \left[ -\frac{\ln(n_p/\alpha_p)^2}{2(\ln \kappa_p)^2} \right]; \quad (37)$$

To get to the posterior for  $\alpha_p$  given  $n_p$  we need an ur-prior  $\eta(\alpha_p)$

$$\pi(\alpha_p) \propto \eta(\alpha_p) \frac{1}{\sqrt{2\pi} \ln \kappa} \frac{1}{n_p} \exp \left[ -\frac{\ln(n_p/\alpha_p)^2}{2(\ln \kappa_p)^2} \right] \quad (38)$$

If  $\eta(\alpha_p)$  is uniform, then the posterior looks like the red curve in Fig. 7(b). However, when paired with an “ur-prior”  $\eta(\alpha_p) \propto 1/\alpha_p$  (green curve in Fig. 7(b)), this results in a posterior distribution that is also of a log-normal form for  $\alpha_p$  (blue curve in Fig. 7(b)).



**Fig. 7:** The lognormal constraint term: (left) the pdf for the global observable  $a_p$  and (right) the likelihood function, the posterior based on a flat prior on  $\alpha_p$ , and the posterior based on a  $1/\alpha_p$  prior.

#### 4.1.7 Incorporating Monte Carlo statistical uncertainty on the histogram templates

The histogram based approach described above are based Monte Carlo simulations of full detector simulation. These simulations are very computationally intensive and often the histograms are sparsely



populated. In this case the histograms are not good descriptions of the underlying distribution, but are estimates of that distribution with some statistical uncertainty. Barlow and Beeston outlined a treatment of this situation in which each bin of each sample is given a nuisance parameter for the true rate, which is then fit using both the data measurement and the Monte Carlo estimate [17]. This approach would lead to several hundred nuisance parameters in the current analysis. Instead, the HistFactory employs a lighter weight version in which there is only one nuisance parameter per bin associated with the total Monte Carlo estimate and the total statistical uncertainty in that bin. If we focus on an individual bin with index  $b$  the contribution to the full statistical model is the factor

$$\text{Pois}(n_b|\nu_b(\boldsymbol{\alpha}) + \gamma_b\nu_b^{\text{MC}}(\boldsymbol{\alpha})) \text{Pois}(m_b|\gamma_b\tau_b), \quad (39)$$

where  $n_b$  is the number of events observed in the bin,  $\nu_b(\boldsymbol{\alpha})$  is the number of events expected in the bin where Monte Carlo statistical uncertainties need not be included (either because the estimate is data driven or because the Monte Carlo sample is sufficiently large),  $\nu_b^{\text{MC}}(\boldsymbol{\alpha})$  is the number of events estimated using Monte Carlo techniques where the statistical uncertainty needs to be taken into account. Both expectations include the dependence on the parameters  $\boldsymbol{\alpha}$ . The factor  $\gamma_b$  is the nuisance parameter reflecting that the true rate may differ from the Monte Carlo estimate  $\nu_b^{\text{MC}}(\boldsymbol{\alpha})$  by some amount. If the total statistical uncertainty is  $\delta_b$ , then the relative statistical uncertainty is given by  $\nu_b^{\text{MC}}/\delta_b$ . This corresponds to a total Monte Carlo sample in that bin of size  $m_b = (\delta_b/\nu_b^{\text{MC}})^2$ . Treating the Monte Carlo estimate as an auxiliary measurement, we arrive at a Poisson constraint term  $\text{Pois}(m_b|\gamma_b\tau_b)$ , where  $m_b$  would fluctuate about  $\gamma_b\tau_b$  if we generated a new Monte Carlo sample. Since we have scaled  $\gamma$  to be a factor about 1, then we also have  $\tau_b = (\nu_b^{\text{MC}}/\delta_b)^2$ ; however,  $\tau_b$  is treated as a fixed constant and does not fluctuate when generating ensembles of pseudo-experiments.

It is worth noting that the conditional maximum likelihood estimate  $\hat{\gamma}_b(\boldsymbol{\alpha})$  can be solved analytically with a simple quadratic expression.

$$\hat{\gamma}_b(\boldsymbol{\alpha}) = \frac{-B + \sqrt{B^2 - 4AC}}{2A}, \quad (40)$$

with

$$A = \nu_b^{\text{MC}}(\boldsymbol{\alpha})^2 + \tau_b\nu_b^{\text{MC}}(\boldsymbol{\alpha}) \quad (41)$$

$$B = \nu_b(\boldsymbol{\alpha})\tau + \nu_b(\boldsymbol{\alpha})\nu_b^{\text{MC}}(\boldsymbol{\alpha}) - n_b\nu_b^{\text{MC}}(\boldsymbol{\alpha}) - m_b\nu_b^{\text{MC}}(\boldsymbol{\alpha}) \quad (42)$$

$$C = m_b\nu_b(\boldsymbol{\alpha}). \quad (43)$$

In a Bayesian technique with a flat prior on  $\gamma_b$ , the posterior distribution is a gamma distribution. Similarly, the distribution of  $\hat{\gamma}_b$  will take on a skew distribution with an envelope similar to the gamma distribution, but with features reflecting the discrete values of  $m_b$ . Because the maximum likelihood estimate of  $\gamma_b$  will also depend on  $n_b$  and  $\hat{\boldsymbol{\alpha}}$ , the features from the discrete values of  $m_b$  will be smeared. This effect will be more noticeable for large statistical uncertainties where  $\tau_b$  is small and the distribution of  $\hat{\gamma}_b$  will have several small peaks. For smaller statistical uncertainties where  $\tau_b$  is large the distribution of  $\hat{\gamma}_b$  will be approximately Gaussian.

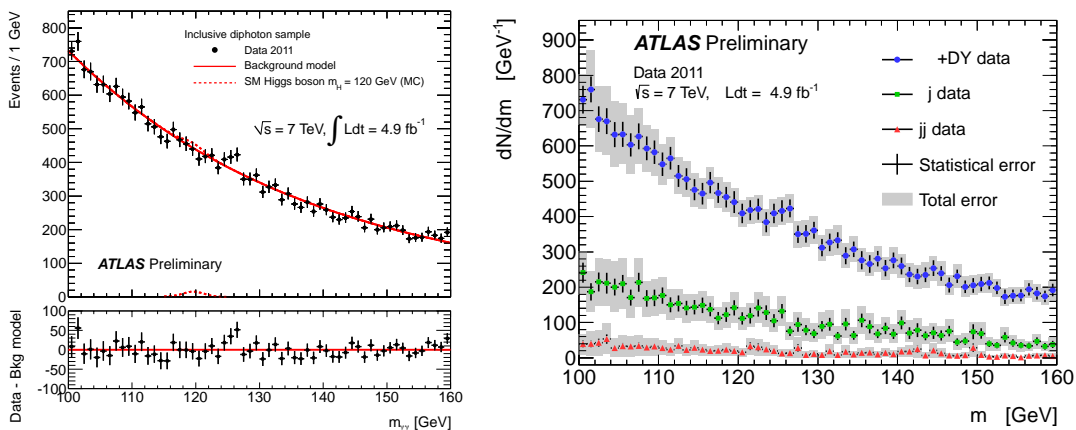
## 4.2 Data-Driven Narrative

The strength of the simulation narrative lies in its direct logical link from the underlying theory to the modeling of the experimental observations. The weakness of the simulation narrative derives from the weaknesses in the simulation itself. Data-driven approaches are more motivated when they address specific deficiencies in the simulation. Before moving to a more abstract or general discussion of the data-driven narrative, let us first consider a few examples.

The first example we have already considered in Sec. 2.2 in the context of the ‘‘on/off’’ problem. There we introduced an auxiliary measurement that counted  $n_{CR}$  events in a control region to estimate

the background  $\nu_B$  in the signal region. In order to do this we needed to understand the ratio of the number of events from the background process in the control and signal regions,  $\tau$ . This ratio  $\tau$  either comes from some reasonable assumption or simulation. For example, if one wanted to estimate the background due to jets faking muons  $j \rightarrow \mu$  for a search selecting  $\mu^+\mu^-$ , then one might use a sample of  $\mu^\pm\mu^\pm$  events as a control region. Here the motivation for using a data-driven approach is that modeling the processes that lead to  $j \rightarrow \mu$  rely heavily on the tails of fragmentation functions and detector response, which one might reasonably have some skepticism. If one assumes that control region is expected to have negligible signal in it, that backgrounds that produce  $\mu^+\mu^-$  other than the jets faking muons, and that the rate for  $j \rightarrow \mu^-$  is the same<sup>18</sup> as the rate for  $j \rightarrow \mu^+$ , then one can assume  $\tau = 1$ . Thus, this background estimate is as trustworthy as the assumptions that went into it. In practice, several of these assumptions may be violated. Another approach is to use simulation of these background processes to estimate the ratio  $\tau$ ; a hybrid of the data-driven and simulation narratives.

Let us now consider the search for  $H \rightarrow \gamma\gamma$  shown in Fig. 8 [18, 19]. The right plot of Fig. 8 shows the composition of the backgrounds in this search, including the continuum production of  $pp \rightarrow \gamma\gamma$ , the  $\gamma$ +jets process with a jet faking a photon  $j \rightarrow \gamma$ , and the multi jet process with two jets faking photons. The continuum production of  $\gamma\gamma$  has a theoretical uncertainty that is much larger than the statistical fluctuations one would expect in the data. Similarly, the rate of jets faking photons is sensitive to fragmentation and the detector simulation. These uncertainties are large compared to the statistical fluctuations in the data itself. Thus we can use the distribution in Fig. 8 to measure the total background rate. Of course, the signal would also be in this distribution, so one either needs to apply a mass window around the signal and consider the region outside of the window as a sideband control sample or model the signal and background contributions to the distribution. In the case of the  $H \rightarrow \gamma\gamma$  shown in Fig. 8 [18, 19] the modeling of the distribution signal and background distributions is not based on histograms from simulation, but instead a continuous function is used as an effective model. I will discuss this effective modeling narrative below, but point out that here this is another example of a hybrid narrative.



**Fig. 8:** Distribution of diphoton invariant mass distributions in the ATLAS  $H \rightarrow \gamma\gamma$  search. The left plot shows a fit of an effective model to the data and the right plot shows an estimate of the  $\gamma\gamma$ ,  $\gamma$ +jet, and dijet contributions.

The final example to consider is an extension of the ‘on/off’ model, often referred to as the ‘ABCD’ method. Let us start with the ‘on/off’ model:  $\text{Pois}(n_{SR}|\nu_S + \nu_B) \cdot \text{Pois}(n_{CR}|\tau\nu_B)$ . As mentioned above, this requires that one estimate  $\tau$  either from simulation or through some assumptions. The ABCD method aims to estimate introduce two new control regions that can be used to measure  $\tau$ . To see this, let us imagine that the signal and control regions correspond to requiring some continuous variable  $x$

<sup>18</sup>Given that the LHC collides  $pp$  and not  $p\bar{p}$ , there is clearly a reason to worry if this assumption is valid.

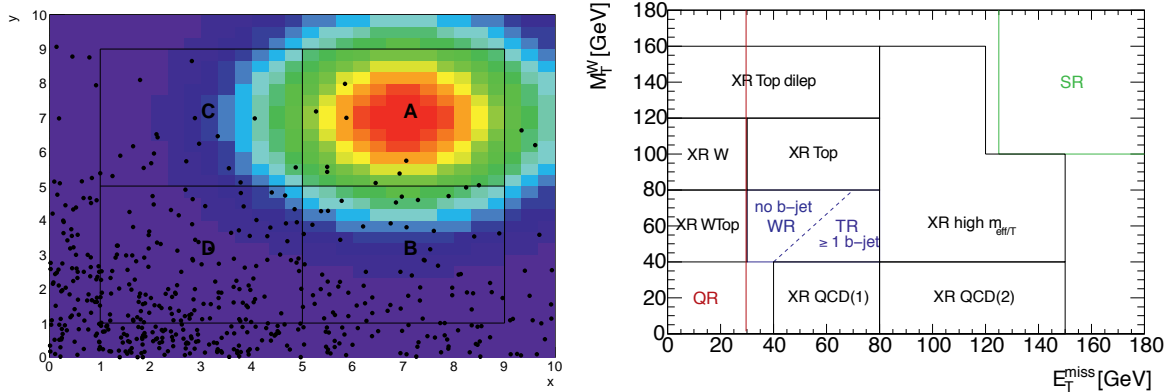
being greater than or less than some threshold value  $x_c$ . If we could introduce a second discriminating variable  $y$  such that the distribution for background factorizes  $f_B(x, y) = f_B(x)f_B(y)$ , then we have a handle to measure the factor  $\tau$ . Typically, one introduces a threshold  $y_c$  so that the signal contribution is small below this threshold<sup>19</sup>. Figure 9 shows an example where  $x_c = y_c = 5$ . With this we these two thresholds we have four regions that we can schematically refer to as A, B, C, and D. In the case of simply counting events in these regions we can write the total expectation as

$$\begin{aligned}
 \nu_A &= 1 \cdot \mu + \nu_A^{MC} + 1 \cdot \nu_A \\
 \nu_B &= \epsilon_B \mu + \nu_B^{MC} + \tau_B \nu_A \\
 \nu_C &= \epsilon_C \mu + \nu_C^{MC} + \tau_C \nu_A \\
 \nu_D &= \epsilon_D \mu + \nu_D^{MC} + \tau_B \tau_C \nu_A
 \end{aligned} \tag{44}$$

where  $\mu$  is the signal rate in region A,  $\epsilon_i$  is the ratio of the signal in the regions B, C, D with respect to the signal in region A,  $\nu_i^{MC}$  is the rate of background in each of the regions being estimated from simulation,  $\nu_i$  is the rate of the background being estimated with the data driven technique in the signal region, and  $\tau_i$  are the ratios of the background rates in the regions B, C, and D with respect to the background in region A. The key is that we have used the factorization  $f_B(x, y) = f_B(x)f_B(y)$  to write  $\tau_D = \tau_B \tau_C$ . The right panel of Fig. 9 shows a more complicated extension of the ABCD method from a recent ATLAS SUSY analysis [20].

An alternative parametrization, which can be more numerically stable is

$$\begin{aligned}
 \nu_A &= 1 \cdot \mu + \nu_A^{MC} + \eta_C \eta_B \nu_D \\
 \nu_B &= \epsilon_B \mu + \nu_B^{MC} + \eta_B \nu_D \\
 \nu_C &= \epsilon_C \mu + \nu_C^{MC} + \eta_C \nu_D \\
 \nu_D &= \epsilon_D \mu + \nu_D^{MC} + 1 \cdot \nu_D
 \end{aligned} \tag{45}$$



**Fig. 9:** An example of ABCD (from Alex Read) in the  $x - y$  plane of two observables  $x$  and  $y$  (left). A more complex example with several regions in the  $M_T^W - E_T^{\text{miss}}$  plane [20].

### 4.3 Effective Model Narrative

In the simulation narrative the model of discriminating variable distributions  $f(x|\alpha)$  is derived from discrete samples of simulated events  $\{x_1, \dots, x_N\}$ . We discussed above how one can use histograms or

<sup>19</sup>The relative sign of the cut is not important, but has been chosen for consistency with Fig 9.

kernel estimation to approximate the underlying distribution and interpolation strategies to incorporate systematic effects. Another approach is to assume some parametric form for the distribution to serve as an effective model. For example, in the  $H \rightarrow \gamma\gamma$  analysis shown in Fig. 8 a simple exponential distribution was used to model the background. The state-of-the-art theoretical predictions for the continuum  $\gamma\gamma$  background process do not predict exactly an exponentially falling distribution, and the analysis must (and does) incorporate the systematic associated to the effective model. Similarly, it is common to use a polynomial in some limited sideband region to estimate backgrounds under a peak. These effective models can range from very ad hoc<sup>20</sup> to more motivated. For instance, one might use knowledge of kinematics and phase space and/or detector resolution to construct an effective model that captures the relevant physics. The advantage of a well motivated effective model is that few nuisance parameters may describe well the relevant family of probability densities, which is the challenge for generic (and relatively unsophisticated) interpolation strategies usually employed in the simulation narrative.

#### 4.4 The Matrix Element Method

Ideally, one would not use a single discriminating variable to distinguish the process of interest from the other background processes, but instead would use as much discriminating power as possible. This implies forming a probability model over a multi-dimensional discriminating variable (ie. a multivariate analysis technique). In principle, both the histogram-based and kernel-based approach generalize to distributions of multi-dimensional discriminating variables; however, in practice, they are limited to only a few dimensions. In the case of histograms this is particularly severe unless one employs clever binning choices, while in the kernel-based approach one can model up to about 5-dimensional distributions with reasonable Monte Carlo sample sizes. In practice, one often uses multivariate algorithms like Neural Networks or boosted decision trees<sup>21</sup> to map the multiple variables into a single discriminating variable. Often these multivariate techniques are seen as somewhat of a black-box. If we restrict ourselves to discriminating variables associated with the kinematics of final state particles (as opposed to the more detailed signature of particles in the detector), then we can often approximate the detailed simulation of the detector with a parametrized detector response. If we denote the kinematic configuration of all the final state particles in the Lorentz invariant phase space as  $\Phi$ , the initial state as  $i$ , the matrix element (potentially averaged over unmeasured spin configurations) as  $\mathcal{M}(i, \Phi)$ , and the probability due to parton density functions for the initial state  $i$  going into the hard scattering as  $f(i)$ , then we can write that the distribution of the, possibly multi-dimensional, discriminating variable  $x$  as

$$f(x) \propto \int d\Phi f(i) |\mathcal{M}(i, \Phi)|^2 W(x|\Phi), \quad (46)$$

where  $W(x|\Phi)$  is referred to as the transfer function of  $x$  given the final state configuration  $\Phi$ . It is natural to think of  $W(x|\Phi)$  as a conditional distribution, but here I let  $W$  encode the efficiency and acceptance so that we have

$$\frac{\sigma_{\text{eff.}}}{\sigma} = \frac{\int dx \int d\Phi |\mathcal{M}(i, \Phi)|^2 W(x|\Phi)}{\int d\Phi |\mathcal{M}(i, \Phi)|^2}. \quad (47)$$

Otherwise, the equation above looks like another application of Bayes's theorem where  $W(x|\Phi)$  plays the role of the pdf/likelihood function and  $\mathcal{M}(i, \Phi)$  plays the role of the prior over the  $\Phi$ . It is worth pointing out that this is a frequentist use of Bayes's theorem since  $d\Phi$  is the Lorentz invariant phase space which explicitly has a measure associated with it.

<sup>20</sup>For instance, the modeling of  $H \rightarrow ZZ^{(*)} \rightarrow 4l$  described in [21] (see Eq. 2 of the corresponding section)

<sup>21</sup>A useful toolkit for high-energy physics is TMVA, which is packaged with ROOT [22].

#### 4.5 Event-by-event resolution, conditional modeling, and Punzi factors

In some cases one would like to provide a distribution for the discriminating variable  $x$  based conditional on some other observable in the event  $y$ :  $f(x|\alpha, y)$ . For instance, one might want to say that the energy resolution for electrons depends on the energy itself through a well-known calorimeter resolution parametrization like  $\sigma(E)/E = A/\sqrt{E} \oplus B$ . These types of conditional distributions can be built in RooFit. A subtle point studied by Punzi is that if  $f(y|\alpha)$  depends on  $\alpha$  the inference on  $\alpha$  can be biased [23]. In particular, if one is trying to estimate the amount of signal in a sample and the distribution of  $y$  for the signal is different than for the background, the estimate of the signal fraction will be biased. This can be remedied by including terms related to  $f(y|\alpha)$ , colloquially called ‘Punzi Factors’. Importantly, this means one cannot build conditional models like this without knowing or assuming something about  $f(y|\alpha)$ .

### 5 Frequentist Statistical Procedures

Here I summarize the procedure used by the LHC Higgs combination group for computing frequentist  $p$ -values uses for quantifying the agreement with the background-only hypothesis and for determining exclusion limits. The procedures are based on the profile likelihood ratio test statistic.

The parameter of interest is the overall signal strength factor  $\mu$ , which acts as a scaling to the total rate of signal events. We often write  $\mu = \sigma/\sigma_{SM}$ , where  $\sigma_{SM}$  is the standard model production cross-section; however, it should be clarified that the same  $\mu$  factor is used for all production modes and could also be seen as a scaling on the branching ratios. The signal strength is called so that  $\mu = 0$  corresponds to the background-only model and  $\mu = 1$  is the standard model signal. It is convenient to separate the full list of parameters  $\alpha$  into the parameter of interest  $\mu$  and the nuisance parameters  $\theta$ :  $\alpha = (\mu, \theta)$ .

For a given data set  $\mathcal{D}_{\text{sim}}$  and values for the global observables  $\mathcal{G}$  there is an associated likelihood function over  $\mu$  and  $\theta$  derived from combined model over all the channels including all the constraint terms in Eq. 6

$$L(\mu, \theta; \mathcal{D}_{\text{sim}}, \mathcal{G}) = \mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G}|\mu, \theta) . \quad (48)$$

The notation  $L(\mu, \theta)$  leaves the dependence on the data implicit, which can lead to confusion. Thus, we will explicitly write the dependence on the data when the identity of the dataset is important and only suppress  $\mathcal{D}_{\text{sim}}, \mathcal{G}$  when the statements about the likelihood are generic.

We begin with the definition of the procedure in the abstract and then describe three implementations of the method based on asymptotic distributions, ensemble tests (Toy Monte Carlo), and importance sampling.

#### 5.1 The test statistics and estimators of $\mu$ and $\theta$

This definitions in this section are all relative to a given dataset  $\mathcal{D}_{\text{sim}}$  and value of the global observables  $\mathcal{G}$ , thus we will suppress their appearance. The nomenclature follows from Ref. [1].

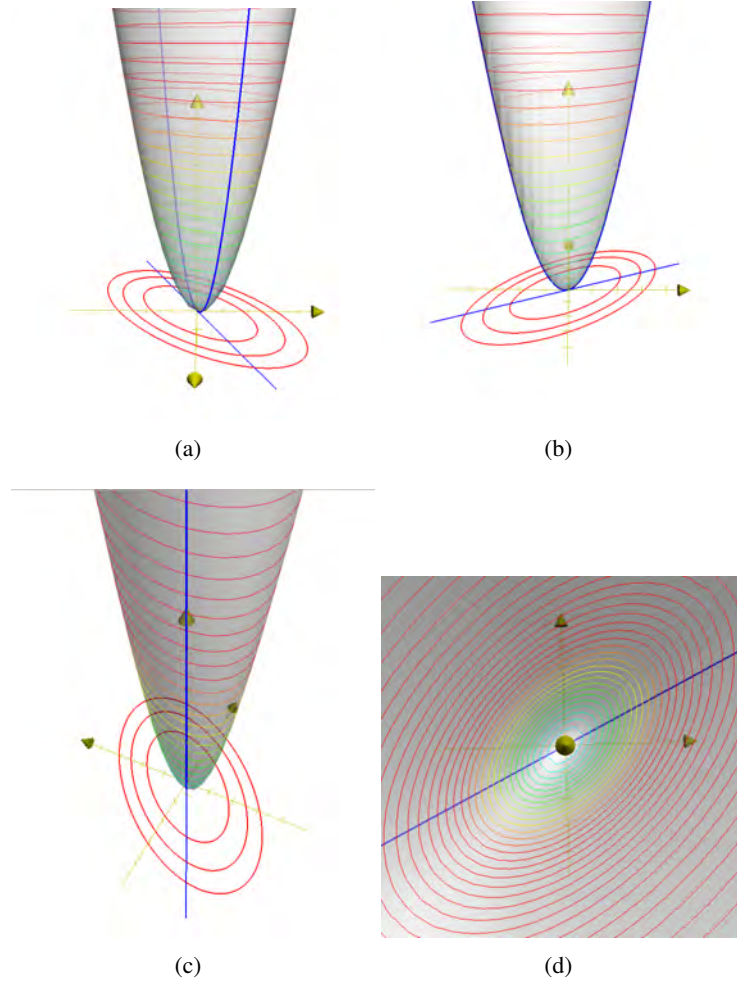
The maximum likelihood estimates (MLEs)  $\hat{\mu}$  and  $\hat{\theta}$  and the values of the parameters that maximize the likelihood function  $L(\mu, \theta)$  or, equivalently, minimize  $-\ln L(\mu, \theta)$ . The dependence of the likelihood function on the data propagates to the values of the MLEs, so when needed the MLEs will be given subscripts to indicate the data set used. For instance,  $\hat{\theta}_{\text{obs}}$  is the MLE of  $\theta$  derived from the observed data and global observables.

The conditional maximum likelihood estimate (CMLEs)  $\hat{\theta}(\mu)$  is the value of  $\theta$  that maximizes the likelihood function with  $\mu$  fixed; it can be seen as a multidimensional function of the single variable  $\mu$ . Again, the dependence on  $\mathcal{D}_{\text{sim}}$  and  $\mathcal{G}$  is implicit. This procedure for choosing specific values of the nuisance parameters for a given value of  $\mu$ ,  $\mathcal{D}_{\text{sim}}$ , and  $\mathcal{G}$  is often referred to as ‘‘profiling’’. Similarly,  $\hat{\theta}(\mu)$  is often called ‘‘the profiled value of  $\theta$ ’’.

Given these definitions, we can construct the profile likelihood ratio

$$\lambda(\mu) = \frac{L(\mu, \hat{\boldsymbol{\theta}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}, \quad (49)$$

which depends explicitly on the parameter of interest  $\mu$ , implicitly on the data  $\mathcal{D}_{\text{sim}}$  and global observables  $\mathcal{G}$ , and is independent of the nuisance parameters  $\boldsymbol{\theta}$  (which have been eliminated via ‘‘profiling’’).



**Fig. 10:** Visualization of a two dimensional likelihood function  $-2 \ln L(\mu, \theta)$ . The blue line in the plane represents the profiling operation  $\hat{\boldsymbol{\theta}}(\mu)$  and the blue curve along the likelihood surface represents  $-2 \ln \lambda(\mu)$ . Note it is was to show that the blue line exits the contours of  $-2 \ln L(\mu, \theta)$  when they are perpendicular to the  $\mu$  axis, which provides the correspondence between the profile likelihood ratio and the description of the Minos algorithm.

In any physical theory the rate of signal events is non-negative, thus  $\mu \geq 0$ . However, it is often convenient to allow  $\mu < 0$  (as long as the pdf  $f_c(x_c|\mu, \boldsymbol{\theta}) \geq 0$  everywhere). In particular,  $\hat{\mu} < 0$  indicates a deficit of events signal-like with respect to the background only and the boundary at  $\mu = 0$  complicates the asymptotic distributions. Ref. [1] uses a trick that is equivalent to requiring  $\mu \geq 0$  while avoiding the formal complications of a boundary, which is to allow  $\mu < 0$  and impose the constraint in the test

statistic itself. In particular, one defines  $\tilde{\lambda}(\mu)$

$$\tilde{\lambda}(\mu) = \begin{cases} \frac{L(\mu, \hat{\boldsymbol{\theta}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} & \hat{\mu} \geq 0, \\ \frac{L(\mu, \hat{\boldsymbol{\theta}}(\mu))}{L(0, \hat{\boldsymbol{\theta}}(0))} & \hat{\mu} < 0 \end{cases} \quad (50)$$

This is not necessary when ensembles of pseudo-experiments are generated with ‘‘Toy’’ Monte Carlo techniques, but since they are equivalent we will write  $\tilde{\lambda}$  to emphasize the boundary at  $\mu = 0$ .

For discovery the test statistic  $\tilde{q}_0$  is used to differentiate the background-only hypothesis  $\mu = 0$  from the alternative hypothesis  $\mu > 0$ :

$$\tilde{q}_0 = \begin{cases} -2 \ln \tilde{\lambda}(\mu) & \hat{\mu} > 0 \\ 0 & \hat{\mu} \leq 0 \end{cases} \quad (51)$$

Note that  $\tilde{q}_0$  is test statistic for a one-sided alternative. Note also that if we consider the parameter of interest  $\mu \geq 0$ , then it is equivalent to the two-sided test (because there are no values of  $\mu$  less than  $\mu = 0$ ).

For limit setting the test statistic  $\tilde{q}_\mu$  is used to differentiate the hypothesis of signal being produced at a rate  $\mu$  from the alternative hypothesis of signal events being produced at a lesser rate  $\mu' < \mu$ :

$$\tilde{q}_\mu = \begin{cases} -2 \ln \tilde{\lambda}(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} = \begin{cases} -2 \ln \frac{L(\mu, \hat{\boldsymbol{\theta}}(\mu))}{L(0, \hat{\boldsymbol{\theta}}(0))} & \hat{\mu} < 0, \\ -2 \ln \frac{L(\mu, \hat{\boldsymbol{\theta}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} & 0 \leq \hat{\mu} \leq \mu, \\ 0 & \hat{\mu} > \mu. \end{cases} \quad (52)$$

Note that  $\tilde{q}_\mu$  is a test statistic for a one-sided alternative; it is a test statistic for a one-sided upper limit.

The test statistic  $\tilde{t}_\mu$  is used to differentiate signal being produced at a rate  $\mu$  from the alternative hypothesis of signal events being produced at a lesser or greater rate  $\mu' \neq \mu$ .

$$\tilde{t}_\mu = -2 \ln \tilde{\lambda}(\mu) . \quad (53)$$

Note that  $\tilde{t}_\mu$  is a test statistic for a two-sided alternative (as in the case of the Feldman-Cousins technique, though this is more general as it incorporates nuisance parameters). Note that if we consider the parameter of interest  $\mu \geq 0$  and we test at  $\mu = 0$  then there is no ‘‘other side’’ and we have  $\tilde{t}_{\mu=0} = \tilde{q}_0$ . Finally, if one relaxes the constraint  $\mu \geq 0$  then the two-sided test statistic is written  $t_\mu$  or, simply,  $-2 \ln \lambda(\mu)$ .

## 5.2 The distribution of the test statistic and $p$ -values

The test statistic should be interpreted as a single real-valued number that represents the outcome of the experiment. More formally, it is a mapping of the data to a single real-valued number:  $\tilde{q}_\mu : \mathcal{D}_{\text{sim}}, \mathcal{G} \rightarrow \mathbb{R}$ . For the observed data the test statistic has a given value, eg.  $\tilde{q}_{\mu, \text{obs}}$ . If one were to repeat the experiment many times the test statistic would take on different values, thus, conceptually, the test statistic has a distribution. Similarly, we can use our model to generate pseudo-experiments using Monte Carlo techniques or more abstractly consider the distribution. Since the number of expected events  $\nu(\mu, \boldsymbol{\theta})$  and the distributions of the discriminating variables  $f_c(x_c | \mu, \boldsymbol{\theta})$  explicitly depend on  $\boldsymbol{\theta}$  the distribution of the test statistic will also depend on  $\boldsymbol{\theta}$ . Let us denote this distribution

$$f(\tilde{q}_\mu | \mu, \boldsymbol{\theta}) , \quad (54)$$

and we have analogous expressions for each of the test statistics described above.

The  $p$ -value for a given observation under a particular hypothesis  $(\mu, \theta)$  is the probability for an equally or more ‘extreme’ outcome than observed assuming that hypothesis

$$p_{\mu, \theta} = \int_{\tilde{q}_{\mu, \text{obs}}}^{\infty} f(\tilde{q}_{\mu} | \mu, \theta) d\tilde{q}_{\mu} . \quad (55)$$

The logic is that small  $p$ -values are evidence against the corresponding hypothesis. In Toy Monte Carlo approaches, the integral above is really carried out in the space of the data  $\int d\mathcal{D}_{\text{sim}} d\mathcal{G}$ .

The immediate difficulty is that we are interested in  $\mu$  but the  $p$ -values depend on both  $\mu$  and  $\theta$ . In the frequentist approach the hypothesis  $\mu = \mu_0$  would not be rejected unless the  $p$ -value is sufficiently small *for all* values of  $\theta$ . Equivalently, one can use the supremum  $p$ -value for over all  $\theta$  to base the decision to accept or reject the hypothesis at  $\mu = \mu_0$ .

$$p_{\mu}^{\text{sup}} = \sup_{\theta} p_{\mu, \theta} \quad (56)$$

The key conceptual reason for choosing the test statistics based on the profile likelihood ratio is that asymptotically (ie. when there are many events) the distribution of the profile likelihood ratio  $\lambda(\mu = \mu_{\text{true}})$  is independent of the values of the nuisance parameters. This follows from Wilks’s theorem. In that limit  $p_{\mu}^{\text{sup}} = p_{\mu, \theta}$  for all  $\theta$ .

The asymptotic distributions  $f(\lambda(\mu) | \mu, \theta)$  and  $f(\lambda(\mu) | \mu', \theta)$  are known and described in Sec. 5.5. For results based on generating ensembles of pseudo-experiments using Toy Monte Carlo techniques does not assume the form of the distribution  $f(\tilde{q}_{\mu} | \mu, \theta)$ , but knowing that it is approximately independent of  $\theta$  means that one does not need to calculate  $p$ -values for all  $\theta$  (which is not computationally feasible). Since there may still be some residual dependence of the  $p$ -values on the choice of  $\theta$  we would like to know the specific value of  $\theta^{\text{sup}}$  that produces the supremum  $p$ -value over  $\theta$ . Since larger  $p$ -values indicate better agreement of the data with the model, it is not surprising that choosing  $\theta^{\text{sup}} = \hat{\theta}(\mu)$  is a good estimate of  $\theta^{\text{sup}}$ . This has been studied in detail by statisticians, and is called the Hybrid Resampling method and is referred to in physics as the ‘profile construction’ [8, 11, 24].

Based on the discussion above, the following  $p$ -value is used to quantify consistency with the hypothesis of a signal strength of  $\mu$ :

$$p_{\mu} = \int_{\tilde{q}_{\mu, \text{obs}}}^{\infty} f(\tilde{q}_{\mu} | \mu, \hat{\theta}(\mu, \text{obs})) d\tilde{q}_{\mu} . \quad (57)$$

A standard 95% confidence-level, one-sided frequentist confidence interval (upper limit) is obtained by solving for  $p'_{\mu_{\text{up}}} = 5\%$ . For downward fluctuations the upper limit of the confidence interval can be arbitrarily small, though it will always include  $\mu = 0$ . This feature is considered undesirable since a physicist would not claim sensitivity to an arbitrarily small signal rate. The feature was the motivation for the modified frequentist method called  $CL_s$  [25–27]. and the alternative approach called power-constrained limits [28].

To calculate the  $CL_s$  upper limit, we define  $p'_{\mu}$  as a ratio of  $p$ -values,

$$p'_{\mu} = \frac{p_{\mu}}{1 - p_b} , \quad (58)$$

where  $p_b$  is the  $p$ -value derived from the same test statistic under the background-only hypothesis

$$p_b = 1 - \int_{\tilde{q}_{\mu, \text{obs}}}^{\infty} f(\tilde{q}_{\mu} | 0, \hat{\theta}(\mu = 0, \text{obs})) d\tilde{q}_{\mu} . \quad (59)$$

The  $CL_s$  upper-limit on  $\mu$  is denoted  $\mu_{\text{up}}$  and obtained by solving for  $p'_{\mu_{\text{up}}} = 5\%$ . It is worth noting that while confidence intervals produced with the ‘‘CLs’’ method over cover, a value of  $\mu$  is regarded



as excluded at the 95% confidence level if  $\mu < \mu_{up}$ . The amount of over coverage is not immediately obvious; however, for small values of  $\mu$  the coverage approaches 100% and for large values of  $\mu$  the coverage is near the nominal 95% (due to  $\langle p_b \rangle \approx 0$ ).

For the purposes discovery one is interested in compatibility of the data with the background-only hypothesis. Statistically, a discovery corresponds to rejecting the background-only hypothesis. This compatibility is based on the following  $p$ -value

$$p_0 = \int_{\tilde{q}_0, obs}^{\infty} f(\tilde{q}_0|0, \hat{\theta}(\mu = 0, obs)) d\tilde{q}_0 . \quad (60)$$

This  $p$ -value is also based on the background-only hypothesis, but the test statistic  $\tilde{q}_0$  is suited for testing the background-only while the test statistic  $\tilde{q}_\mu$  in Eq. 59 is suited for testing a hypothesis with signal.

It is customary to convert the background-only  $p$ -value into the quantile (or ‘‘sigma’’) of a unit Gaussian. This conversion is purely conventional and makes no assumption that the test statistic  $q_0$  is Gaussian distributed. The conversion is defined as:

$$Z = \Phi^{-1}(1 - p_0); \quad (61)$$

where  $\Phi^{-1}$  is the inverse of the cumulative distribution for a unit Gaussian. One says the significance of the result is  $Z\sigma$  and the standard discovery convention is  $5\sigma$ , corresponding to  $p_0 = 2.87 \cdot 10^{-7}$ .

### 5.3 Expected sensitivity and bands

The expected sensitivity for limits and discovery are useful quantities, though subject to some degree of ambiguity. Intuitively, the expected upper limit is the upper limit one would expect to obtain if the background-only hypothesis is true. Similarly, the expected significance is the significance of the observation assuming the standard model signal rate (at some  $m_H$ ). To find the expected limit one needs a distribution  $f(\mu_{up}|\mu = 0, \theta)$ . To find the expected significance one needs the distribution  $f(Z|\mu = 1, \theta)$  or, equivalently,  $f(p_0|\mu = 1, \theta)$ . We use the median instead of the mean, as it is invariant to the choice of  $Z$  or  $p_0$ . More importantly, is that the expected limit and significance depend on the value of the nuisance parameters  $\theta$ , for which we do not know the true values. Thus, the expected limit and significance will depend on some convention for choosing  $\theta$ . While many nuisance parameters have a nominal estimate (i.e. the global observables in the constraint terms), others do not (eg. the exponent in the  $H \rightarrow \gamma\gamma$  background model). Thus, we choose a convention that treats all of the nuisance parameters consistently, which is the profiled value based on the observed data. Thus for the expected limit we use  $f(\mu_{up}|0, \hat{\theta}(\mu = 0, obs))$  and for the expected significance we use  $f(p_0|\mu = 1, \hat{\theta}(\mu = 1, obs))$ . An unintuitive and possibly undesirable feature of this choice is that the expected limit and significance depend on the observed data through the conventional choice for  $\theta$ .

With these distributions we can also define bands around the median upper limit. Our standard limit plot shows a dark green band corresponding to  $\mu_{\pm 1}$  defined by

$$\int_0^{\mu_{\pm 1}} f(\mu_{up}|0, \hat{\theta}(\mu = 0, obs)) d\mu_{up} = \Phi^{-1}(\pm 1) \quad (62)$$

and a light yellow band corresponding to  $\mu_{\pm 2}$  defined by

$$\int_0^{\mu_{\pm 2}} f(\mu_{up}|0, \hat{\theta}(\mu = 0, obs)) d\mu_{up} = \Phi^{-1}(\pm 2) \quad (63)$$

### 5.4 Ensemble of pseudo-experiments generated with ‘‘Toy’’ Monte Carlo

The  $p$ -values in the procedure described above require performing several integrals. In the case of the asymptotic approach, the distributions for  $\tilde{q}_\mu$  and  $\tilde{q}_0$  are known and the integral is performed directly.

When the distributions are not assumed to take on their asymptotic form, then they must be constructed using Monte Carlo methods. In the “toy Monte Carlo” approach one generates pseudo-experiments in which the number of events in each channel  $n_c$ , the values of the discriminating variables  $\{x_{ec}\}$  for each of those events, and the auxiliary measurements (global observables)  $a_p$  are all randomized according to  $\mathbf{f}_{\text{tot}}$ . We denote the resulting data  $\mathcal{D}_{\text{toy}}$  and global observables  $\mathcal{G}_{\text{toy}}$ . By doing this several times one can build an ensemble of pseudo-experiments and evaluate the necessary integrals. Recall that Monte Carlo techniques can be viewed as a form of numerical integration.

The fact that the auxiliary measurements  $a_p$  are randomized is unfamiliar in particle physics. The more familiar approach for toy Monte Carlo is that the nuisance parameters are randomized. This requires a distribution for the nuisance parameters, and thus corresponds to a Bayesian treatment of the nuisance parameters. The resulting  $p$ -values are a hybrid Bayesian-Frequentist quantity with no consistent definition of probability. To maintain a strictly frequentist procedure, the corresponding operation is to randomize the auxiliary measurements.

While formally this procedure is well motivated, as physicists we also know that our models can have deficiencies and we should check that the distribution of the auxiliary measurements does not deviate too far from our expectations.

Technically, the pseudo-experiments are generated with the RooStats `ToyMCSampler`, which is used by the higher-level tool `FrequentistCalculator`, which is in turn used by `HypoTestInverter`.

## 5.5 Asymptotic Formulas

The following has been extracted from Ref. [1] and has been reproduced here for convenience. The primary message of Ref. [1] is that for a sufficiently large data sample the distributions of the likelihood ratio based test statistics above converge to a specific form. In particular, Wilks’s theorem [29] can be used to obtain the distribution  $f(\lambda(\mu)|\mu)$ , that is the distribution of the test statistic  $\lambda(\mu)$  when  $\mu$  is true. Note that the asymptotic distribution is independent of the value of the nuisance parameters. Wald’s theorem [30] provides the generalization to  $f(\lambda(\mu)|\mu', \boldsymbol{\theta})$ , that is when the true value is not the same as the tested value. The various formulae listed below are corollaries of Wilks’s and Wald’s theorems for the likelihood ratio test statistics described above. The Asimov data described immediately below was a novel result of Ref. [1].

### 5.5.1 The Asimov data and $\sigma = \text{var}(\hat{\mu})$

The asymptotic formulae below require knowing the variance of the maximum likelihood estimate of  $\mu$

$$\sigma = \text{var}[\hat{\mu}] . \quad (64)$$

One result of Ref. [1] is that  $\sigma$  can be estimated with an artificial dataset referred to as the *Asimov* dataset. The Asimov dataset is defined as a binned dataset, where the number of events in bin  $b$  is exactly the number of events expected in bin  $b$ . Note, this means that the dataset generally has non-integer number of events in each bin. For our general model one can write

$$n_{b,A} = \int_{x \in \text{bin } b} \nu(\boldsymbol{\alpha}) f(x|\boldsymbol{\alpha}) dx \quad (65)$$

where the subscript  $A$  denotes that this is the Asimov data. Note, that the dataset depends on the value of  $\boldsymbol{\alpha}$  implicitly. For an model of unbinned data, one can simply take the limit of narrow bin widths for the Asimov data. We denote the likelihood evaluated with the Asimov data as  $L_A(\mu)$ . The important result is that one can calculate the expected Fisher information of Eq. 7 by computing the observed Fisher information on the likelihood function based on this special Asimov dataset.

A related and convenient way to calculate the variance of  $\hat{\mu}$  is

$$\sigma \sim \frac{\mu}{\sqrt{\tilde{q}_{\mu,A}}} . \quad (66)$$

where  $\tilde{q}_{\mu,A}$  is the to use the  $\tilde{q}_\mu$  test statistic based on a background-only Asimov data (ie. the one with  $\mu = 0$  in Eq. 65). It is worth noting that higher-order corrections to the formulae below are being developed to address the case when the variance of  $\hat{\mu}$  depends strongly on  $\mu$ .

### 5.5.2 Asymptotic Formulas for $\tilde{q}_0$

For a sufficiently large data sample, the pdf  $f(\tilde{q}_0|\mu')$  is found to approach

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right) \delta(q_0) + \frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{q_0}} \exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]. \quad (67)$$

For the special case of  $\mu' = 0$ , this reduces to

$$f(q_0|0) = \frac{1}{2} \delta(q_0) + \frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{q_0}} e^{-q_0/2}. \quad (68)$$

That is, one finds a mixture of a delta function at zero and a chi-square distribution for one degree of freedom, with each term having a weight of 1/2. In the following we will refer to this mixture as a half chi-square distribution or  $\frac{1}{2}\chi_1^2$ .

From Eq. (67) the corresponding cumulative distribution is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right). \quad (69)$$

The important special case  $\mu' = 0$  is therefore simply

$$F(q_0|0) = \Phi\left(\sqrt{q_0}\right). \quad (70)$$

The  $p$ -value of the  $\mu = 0$  hypothesis is

$$p_0 = 1 - F(q_0|0), \quad (71)$$

and therefore for the significance gives the simple formula

$$Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}. \quad (72)$$

### 5.5.3 Asymptotic Formulas for $\tilde{q}_\mu$

For a sufficiently large data sample, the pdf  $f(\tilde{q}_\mu|\mu)$  is found to approach

$$f(\tilde{q}_\mu|\mu) = \Phi\left(\frac{\mu' - \mu}{\sigma}\right) \delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\tilde{q}_\mu}} \exp\left[-\frac{1}{2}\left(\sqrt{\tilde{q}_\mu} - \frac{\mu - \mu'}{\sigma}\right)^2\right] & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2 \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu - (\mu^2 - 2\mu\mu')/\sigma^2)^2}{(2\mu/\sigma)^2}\right] & \tilde{q}_\mu > \mu^2/\sigma^2 \end{cases}. \quad (73)$$

The special case  $\mu = \mu'$  is therefore

$$f(\tilde{q}_\mu|\mu) = \frac{1}{2} \delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\tilde{q}_\mu}} e^{-\tilde{q}_\mu/2} & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2 \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu + \mu^2/\sigma^2)^2}{(2\mu/\sigma)^2}\right] & \tilde{q}_\mu > \mu^2/\sigma^2. \end{cases} \quad (74)$$

The corresponding cumulative distribution is

$$F(\tilde{q}_\mu|\mu') = \begin{cases} \Phi\left(\sqrt{\tilde{q}_\mu} - \frac{\mu - \mu'}{\sigma}\right) & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2, \\ \Phi\left(\frac{\tilde{q}_\mu - (\mu^2 - 2\mu\mu')/\sigma^2}{2\mu/\sigma}\right) & \tilde{q}_\mu > \mu^2/\sigma^2. \end{cases} \quad (75)$$

The special case  $\mu = \mu'$  is

$$F(\tilde{q}_\mu|\mu) = \begin{cases} \Phi\left(\sqrt{\tilde{q}_\mu}\right) & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2, \\ \Phi\left(\frac{\tilde{q}_\mu + \mu^2/\sigma^2}{2\mu/\sigma}\right) & \tilde{q}_\mu > \mu^2/\sigma^2. \end{cases} \quad (76)$$

The  $p$ -value of the hypothesized  $\mu$  is as before given by one minus the cumulative distribution,

$$p_\mu = 1 - F(\tilde{q}_\mu|\mu). \quad (77)$$

As when using  $q_\mu$ , the upper limit on  $\mu$  at confidence level  $1 - \alpha$  is found by setting  $p_\mu = \alpha$  and solving for  $\mu$ , which reduces to the same result as found when using  $q_\mu$ , namely,

$$\mu_{up} = \hat{\mu} + \sigma\Phi^{-1}(1 - \alpha). \quad (78)$$

Note that because  $\sigma$  depends in general on  $\mu$ , Eq. (78) must be solved numerically.

#### 5.5.4 Expected $CL_s$ Limit and Bands

For the  $CL_s$  method we need distributions for  $\tilde{q}_\mu$  for the hypothesis at  $\mu$  and  $\mu = 0$ . We find

$$p'_\mu = \frac{1 - \Phi(\sqrt{q_\mu})}{\Phi(\sqrt{q_{\mu,A}} - \sqrt{q_\mu})} \quad (79)$$

The median and expected error bands will therefore be

$$\mu_{up+N} = \sigma(\Phi^{-1}(1 - \alpha\Phi(N)) + N) \quad (80)$$

with

$$\sigma^2 = \frac{\mu^2}{q_{\mu,A}} \quad (81)$$

$\alpha = 0.05$ ,  $\mu$  can be taken as  $\mu_{up}^{med}$  in the calculation of  $\sigma$ . Note that for  $N = 0$  we find the median limit

$$\mu_{up}^{med} = \sigma\Phi^{-1}(1 - 0.5\alpha) \quad (82)$$

The fact that  $\sigma$  (the variance of  $\hat{\mu}$ ) defined in Eq. 66 in general depends on  $\mu$  complicates situations and can lead to some discrepancies between the correct value of the bands and those obtained with the equation above. The bands tend to be too narrow. A modified treatment of the bands taking into account the  $\mu$  dependence of  $\sigma$  is under development.

## 5.6 Importance Sampling

[The following section has been adapted from text written primarily by Sven Kreiss, Alex Read, and myself for the ATLAS Higgs combination. It is reproduced here for convenience. ]

To claim a discovery, it is necessary to populate a small tail of a test statistic distribution. Toy Monte-Carlo techniques use the model  $\mathbf{f}_{\text{tot}}$  to generate toy data  $\mathcal{D}_{\text{toy}}$ . For every pseudo-experiment (toy), the test statistic is calculated and added to the test statistic distribution. Building this distribution from toys is independent of the assumptions that go into the asymptotic calculation that describes this distribution with an analytic expression. Recently progress has been made using Importance Sampling to populate the extreme tails of the test statistic distribution, which is much more computationally intensive with standard methods. The presented algorithms are implemented in RooStats ToyMCSampler.

### 5.6.1 Naive Importance Sampling

An ensemble of "standard toys" is generated from a model representing the Null hypothesis with  $\mu = 0$  and the nuisance parameters  $\theta$  fixed at their profiled values to the observed data  $\theta_{\text{obs}}$ , written  $\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G} | \mu = 0, \theta_{\text{obs}})$ . With importance sampling however, the underlying idea is to generate toys from a different model, called the importance density. A valid importance density is for example the same model with a non-zero value of  $\mu$ . The simple Likelihood ratio is calculated for each toy and used as a weight.

$$\text{weight} = \frac{\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{toy}}, \mathcal{G}_{\text{toy}} | \mu = 0, \theta_{\text{obs}})}{\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{toy}}, \mathcal{G}_{\text{toy}} | \mu = \mu', \theta_{\text{obs}})}$$

The weighted distribution is equal to a distribution of unweighted toys generated from the Null. The choice of the importance density is a delicate issue. Michael Woodroffe presented a prescription for creating a well behaved importance density [31]. Unfortunately, this method is impractical for models as large as the combined Higgs models. An alternative approach is shown below.

### 5.6.2 Phase Space Slicing

The first improvement from naive importance sampling is the idea of taking toys from both, the null density and the importance density. There are various ways to do that. Simply stitching two test statistic distributions together at an arbitrary point has the disadvantage that the normalizations of both distributions have to be known.

Instead, it is possible to select toys according to their weights. First, toys are generated from the Null and the simple Likelihood ratio is calculated. If it is larger than one, the toy is kept and otherwise rejected. Next, toys from the importance density are generated. Here again, the simple Likelihood ratio is calculated but this time the toy is rejected when the Likelihood ratio is larger than one and kept when it is smaller than one. If kept, the toy's weight is the simple Likelihood ratio which is smaller than one by this prescription.

In the following section, this idea is restated such that it generalizes to multiple importance densities.

### 5.6.3 Multiple Importance Densities

The above procedure for selecting and reweighting toys that were generated from both densities can be phrased in the following way:

- A toy is generated from a density with  $\mu = \mu'$  and the Likelihoods  $\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{toy}}, \mathcal{G}_{\text{toy}} | \mu = 0, \theta_{\text{obs}})$  and  $\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{toy}}, \mathcal{G}_{\text{toy}} | \mu = \mu', \theta_{\text{obs}})$  are calculated.
- The toy is veto-ed when the Likelihood with  $\mu = \mu'$  is not the largest. Otherwise, the toy is used with a weight that is the ratio of the Likelihoods.

This can be generalized to any number of densities with  $\mu_i = \{0, \mu', \mu'', \dots\}$ . For the toys generated from model  $i$ :

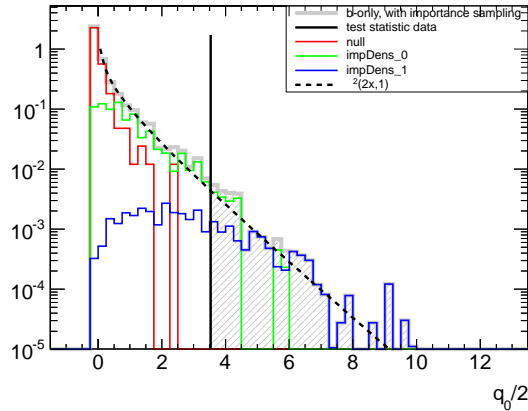
$$\text{veto: if } \mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{toy}}, \mathcal{G}_{\text{toy}} | \mu = \mu_i, \boldsymbol{\theta}_{\text{obs}}) \neq \max \{ \mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{toy}}, \mathcal{G}_{\text{toy}} | \mu = \mu_j, \boldsymbol{\theta}_{\text{obs}}) : \mu_j = \{0, \mu', \mu'', \dots\} \} \quad (83)$$

$$\text{weight} = \frac{\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{toy}}, \mathcal{G}_{\text{toy}} | \mu = 0, \boldsymbol{\theta}_{\text{obs}})}{\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{toy}}, \mathcal{G}_{\text{toy}} | \mu = \mu_i, \boldsymbol{\theta}_{\text{obs}})} \quad (84)$$

The number of importance densities has to be known when applying the vetos. It should not be too small to cover the parameter space appropriately and it should not be too large, because too many importance densities lead to too many vetoed toys which decreases overall efficiency. The value and error of  $\hat{\mu}$  from a fit to data can be used to estimate the required number of importance densities for a given target overlap of the distributions.

The sampling efficiency in the tail can be further improved by generating a larger number of toys for densities with larger values of  $\mu$ . For example, for  $n$  densities, one can generate  $2^k / 2^n = 2^{k-n}$  of the overall toys per density  $k$  with  $k = 0, \dots, n-1$ . The toys have to be re-weighted for example by  $2^{n-1} / 2^k$  resulting in a minimum re-weight factor of one. The current implementation of the error calculation for the p-value is independent of an overall scale in the weights.

The method using multiple importance densities is similar to Michael Woodroffe's [31] prescription of creating a suitable importance density with an integral over  $\mu$ . In the method presented here, the integral is approximated by a sum over discrete values of  $\mu$ . Instead of taking the sum, a mechanism that allows for multiple importance densities is introduced.



**Fig. 11:** An example sampling of a test statistic distribution using three densities, the original null density and two importance densities.

### 5.7 Look-elsewhere effect, trials factor, Bonferoni

Future versions of this document will discuss the so-called look-elsewhere effect in more detail. Here we point to the primary development recently: [32, 33].

### 5.8 One-sided intervals, CLs, power-constraints, and Negatively Biased Relevant Subsets

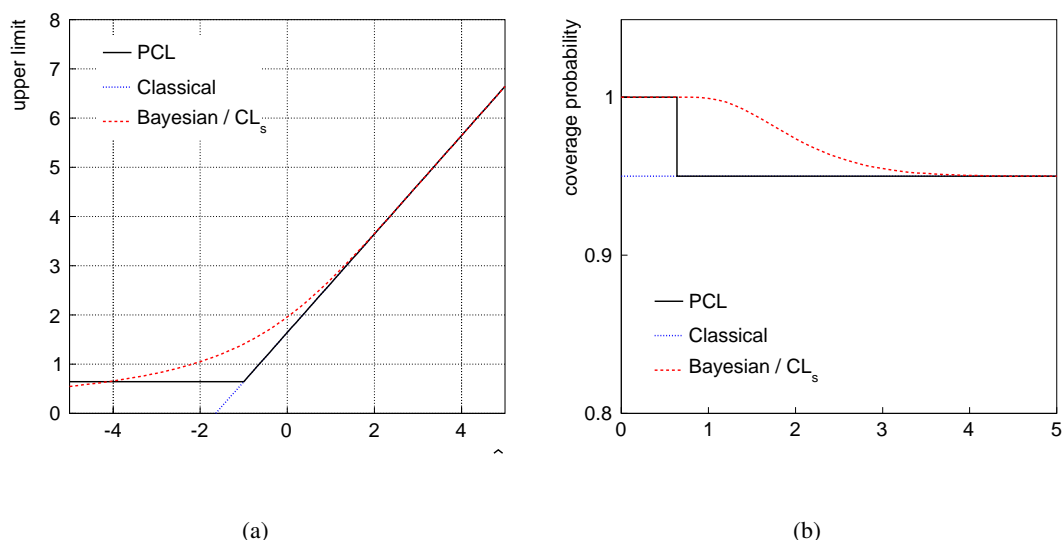
Particle physicists regularly set upper-limits on cross sections and other parameters that are bounded to be non-negative. Standard frequentist confidence intervals should nominally cover at the stated value. The implication that a 95% confidence level upper-limit covers the true value 95% of the time is that it

doesn't cover the true value 5% of the time. This is true no matter how small the cross section is. That means that if there is no signal present, 5% of the time we would be excluding any positive value of the cross-section. Experimentalists do not like this since we would not consider ourselves sensitive to arbitrarily small signals.

Two main approaches have been proposed to protect from excluding signals to which we do not consider ourselves sensitive. The first is the CLs procedure introduced by Read and described above [25–27]. The CLs procedure produce intervals that over-cover – meaning that the intervals cover the true value more than the desired level. The coverage for small values of the cross-section approaches 100%, while for large values of the cross section, where the experiment does have sensitivity, the coverage converges to the nominal level (see Fig. 12). Unfortunately, the coverage for intermediate values is not immediately accessible without more detailed studies. Interestingly, the modified frequentist CLs procedure reproduces the one-sided upper limit from a Bayesian procedure with a uniform prior on the cross section for simple models like number counting analyses. Even in very complicated models we see very good numerical agreement between CLs and the Bayesian approach, even though the interpretation of the numbers is different.

An alternate approach called power-constrained limits (PCL) is to leave the standard frequentist procedure unchanged while adding an additional requirement for a parameter point to be considered 'excluded'. The additional requirement is directly a measure of the sensitivity of to that parameter point based on the notion of power (or Type II error). This approach makes the coverage of the procedure manifest [28].

Surprisingly, one-sided upper limits on a bounded parameter are a subtle topic that has led to debates among the experts of statistics in the collaborations and a string of interesting articles from statisticians. The discussion is beyond the scope of the current version of these notes, but the interested reader is invited and encouraged to read [34] and the responses from notable statisticians on the topic. More recently Cousins tried to formalize the sensitivity problem in terms of a concept called Negatively Biased Relevant Subsets (NBRS) [35]. While the power-constrained limits do not formally emit NBRS, it is an interesting insight. Even more recently, Vitells has found interesting connections with CLs and the work of Birnbaum [27, 36]. This connection is significant since statisticians have primarily seen CLs as an ad hoc procedure mixing the notion of size and power with no satisfying properties.



**Fig. 12:** Taken from Fig.3 of [28]: (a) Upper limits from the PCL (solid), CLs and Bayesian (dashed), and classical (dotted) procedures as a function of  $\mu$ , which is assumed to follow a Gaussian distribution with unit standard deviation. (b) The corresponding coverage probabilities as a function of  $\mu$ .

## 6 Bayesian Procedures

[This section is far from complete. Some key practical issues and references to other literature are given.]

Unsurprisingly, Bayesian procedures are based on Bayes’s theorem as in Eq. 3 and Eq. 5. The Bayesian approach requires one to provide a prior over the parameters, which can be seen either as an advantage or a disadvantage [37, 38]. In practical terms, one typically wants to build the posterior distribution for the parameter of interest. This typically requires integrating, or *marginalizing*, over all the nuisance parameters as in Eq. 14. These integrals can be over very high dimensional posteriors with complicated structure. One of the most powerful algorithms for this integration is Markov Chain Monte Carlo, described below. In terms of the prior one can either embrace the subjective Bayesian approach [39] or take a more ‘objective’ approach in which the prior is derived from formal rules. For instance, Jeffreys’s Prior [40] or their generalization in terms of Reference Priors [41].

Given the logical importance of the choice of prior, it is generally recommended to try a few options to see how the result numerically depends on the choice of priors (i.e.. sensitivity analysis). This leads me to a few great quotes from prominent statisticians:

“Sensitivity analysis is at the heart of scientific Bayesianism” –Michael Goldstein

“Perhaps the most important general lesson is that the facile use of what appear to be uninformative priors is a dangerous practice in high dimensions” -Brad Efron

“Meaningful prior specification of beliefs in probabilistic form over very large possibility spaces is very difficult and may lead to a lot of arbitrariness in the specification” – Michael Goldstein

“Objective Bayesian analysis is the best frequentist tool around” –Jim Berger

### 6.1 Hybrid Bayesian-Frequentist methods

It is worth mentioning that in particle physics there has been widespread use of a hybrid Bayesian-Frequentist approach in which one marginalizes nuisance parameters. Perhaps the most well known example is due to a paper by Cousins and Highland [42]. In some instances one obtains a Bayesian-averaged model that depends only on the parameters of interest

$$\bar{\mathbf{f}}(\mathcal{D}|\alpha_{\text{poi}}) = \int \mathbf{f}_{\text{tot}}(\mathcal{D}|\alpha)\eta(\alpha_{\text{nuis}}) d\alpha_{\text{nuis}} \quad (85)$$

and then proceeds with the typical frequentist methodology for calculating p-values and constructing confidence intervals. Note, in this approach the constraint terms that are appended to  $\mathbf{f}_{\text{sim}}$  of Eq. 2 to obtain  $\mathbf{f}_{\text{tot}}$  of Eq. 6 are interpreted as in Eq. 5 and  $\eta(\alpha_{\text{nuis}})$  is usually a uniform prior. Furthermore, the global observables or auxiliary measurements  $a_p$  are typically left fixed to their nominal or observed values and not randomized. In other variants the full model without constraints  $\mathbf{f}_{\text{sim}}(\mathcal{D}|\alpha)$  is used to define the test statistic but the distribution of the test statistic is obtained by marginalizing (or randomizing) the nuisance parameters as in Eq. 5. See the following references for more details [4, 43–49].

The shortcomings of this approach are that the coverage is not guaranteed and the method uses an inconsistent notion of probability. Thus it is hard to define exactly what the p-values and intervals mean in a formal sense.

### 6.2 Markov Chain Monte Carlo and the Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is used to construct a Markov chain  $\{\alpha_i\}$ , where the samples  $\alpha_i$  are proportional to the target posterior density or likelihood function. The algorithm requires a proposal function  $Q(\alpha|\alpha')$  that gives the probability density to propose the point  $\alpha$  given that the last point in the chain is  $\alpha'$ . Note, the density only depends on the last step in the chain, thus it is considered a Markov process. At each step in the algorithm, a new point in parameter space is proposed and possibly appended to the chain based on its likelihood relative to the current point in the chain. Even when



the proposal density function is not symmetric, Metropolis Hastings maintains ‘detailed balance’ when constructing the Markov chain by counterbalancing the relative likelihood between the two points with the relative proposal density. That is, given the current point  $\alpha$ , proposed point  $\alpha'$ , likelihood function  $L$ , and proposal density function  $Q$ , we visit  $\alpha'$  if and only if

$$\frac{L(\alpha') Q(\alpha|\alpha')}{L(\alpha) Q(\alpha'|\alpha)} \geq \text{Rand}[0, 1] \quad (86)$$

Note, if the proposal density is symmetric,  $Q(\alpha|\alpha') = Q(\alpha'|\alpha)$ , then the ratio of the proposal densities can be neglected (which can be computationally expensive). Above we have written the algorithm to sample the likelihood function  $L(\alpha)$ , but typically one would use the posterior  $\pi(\alpha)$ . Within `RooStats` the Metropolis-Hastings algorithm is implemented with the `MetropolisHastings` class, which returns a `MarkovChain`. Another powerful tool is the Bayesian Analysis Toolkit (BAT) [50]. Note, one can use a `RooFit / RooStats` model in the BAT environment.

Note, an alternative to Markov Chain Monte Carlo is the nested sampling approach of Skilling [51] and the `MultiNest` implementation [52].

Lastly, we mention that sampling algorithms associated to Bayesian belief networks and graphical models may offer enormous advantages to both MCMC and nested sampling due to the fact that they can take advantage of the conditional dependencies in the model.

### 6.3 Jeffreys’s and Reference Prior

One of the great advances in Bayesian methodology was the introduction of Jeffreys’s rule for selecting a prior based on a formal rule [40]. The rule selects a prior that is invariant under reparametrization of the observables and covariant with reparametrization of the parameters. The rule is based on information theoretic arguments and the prior is given by the square root of the determinant of the Fisher information matrix, which we first encountered in Eq. 7.

$$\pi(\alpha) = \sqrt{\det \Sigma_{pp'}^{-1}(\alpha)} = \sqrt{\det \left[ \int \mathbf{f}_{\text{tot}}(\mathcal{D}|\alpha) \frac{-\partial^2 \log \mathbf{f}_{\text{tot}}(\mathcal{D}|\alpha)}{\partial \alpha_p \alpha_{p'}} d\mathcal{D} \right]} \quad (87)$$

While the right-most form of the prior looks daunting with complex integrals over partial derivatives, the Asimov data described in Sec. 5.5.1 and Ref. [1] provide a convenient way to calculate the Fisher information. Fig. 13 and 14 show examples of `RooStats` numerical algorithm for calculating Jeffreys’s prior compared to analytic results on a simple Gaussian and a Poisson model.

Unfortunately, Jeffreys’s prior does not behave well in multidimensional problems. Based on a similar information theoretic approach, Bernardo and Berger have developed the Reference priors [53–56] and the associated Reference analysis. While attractive in many ways, the approach is fairly difficult to implement. Recently, there has been some progress within the particle physics context in deriving the reference prior for problems relevant to particle physics [41, 57].

### 6.4 Likelihood Principle

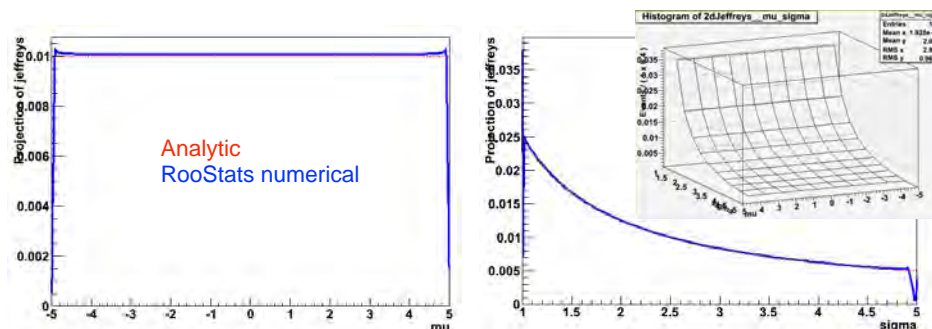
For those interested in the deeper and more philosophical aspects of statistical inference, the likelihood principle is incredibly interesting. This section will be expanded in the future, but for now I simply suggest searching on the internet, the Wikipedia article, and Ref. [36]. In short the principle says that all inference should be based on the likelihood function of the observed data. Frequentist procedures violate the likelihood principle since p-values are tail probabilities associated to hypothetical outcomes (not the observed data). Generally, Bayesian procedures and those based on the asymptotic properties of likelihood tests do obey the likelihood principle. Somewhat ironically, the objective Bayesian procedures such as Reference priors and Jeffreys’s prior can violate the likelihood principle since the prior is based on expectations over hypothetical outcomes.

```

RooWorkspace w("w");
w.factory("Gaussian::g(x[0,-20,20],mu[0,-5,5],sigma[1,0,10])");
w.factory("n[10,.1,200]");
w.factory("ExtendPdf::p(g,n)");
w.var("n")->setConstant();

w.var("sigma")->setConstant();
w.defineSet("poi","mu");
w.defineSet("obs","x");
RooJeffreysPrior pi("jeffreys","jeffreys",*w.pdf("p"),*w.set("poi"),*w.set("obs"));

```



**Fig. 13:** Example code making a Gaussian distribution (with 10 events expected) and the Jeffreys Prior for  $\mu$  and  $\sigma$  calculated numerically in RooStats and compared to the analytic result.

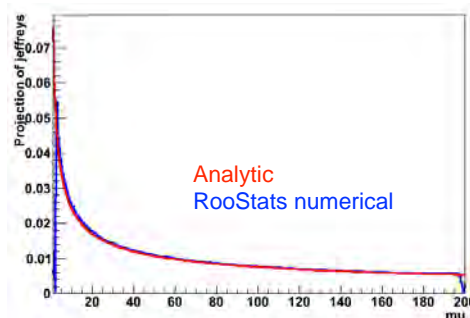
```

RooWorkspace w("w");
w.factory("Uniform::u(x[0,1])");
w.factory("mu[100,1,200]");
w.factory("ExtendPdf::p(u,mu)");

w.defineSet("poi","mu");
w.defineSet("obs","x");
// w.defineSet("obs2","n");

RooJeffreysPrior pi("jeffreys","jeffreys",*w.pdf("p"),*w.set("poi"),*w.set("obs"));

```



**Fig. 14:** Example code making a Poisson distribution (with 100 replications expected) and the Jeffreys Prior for  $\mu$  calculated numerically in RooStats and compared to the analytic result.

## 7 Unfolding

Another topic for the future. The basic aim of unfolding is to try to correct distributions back to the true underlying distribution before detector 'smearing'. For now, see [58–65].

## 8 Conclusions

It was a pleasure to lecture at the 2011 ESHEP school in Cheile Gradistei and the 2013 CLASHEP school in Peru. Quite a bit of progress has been made in the last few years in terms of statistical methodology, in particular the formalization of a fully frequentist approach to incorporating systematics, a deeper understanding of the look-elsewhere effect, the development of asymptotic approximations of the distributions important for particle physics, and in roads to Bayesian reference analysis. Furthermore, most of these developments are general purpose and can be applied across diverse models. While those developments are interesting, the most important area for most physicists to devote their attention in terms of statistics is to improve the modeling of the data for his or her individual analysis.

## References

- [1] G. Cowan, K. Cranmer, E. Gross, O. Vitells. Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J.*, C71:1554, 2011.
- [2] K Cranmer and G. Lewis. The histfactory users guide. <https://twiki.cern.ch/twiki/pub/RooStats>, 2011.
- [3] R. Barlow. Extended maximum likelihood. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 297(3):496 – 506, 1990.
- [4] R. D. Cousins, J. T. Linnemann, and J. Tucker. Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process. *Nucl. Instrum. Meth.*, A595:480–501, 2008.
- [5] F. James and M. Roos. Errors on ratios of small numbers of events. *Nuclear Physics B*, 172:475, 1980.
- [6] F. James and M. Roos. Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations. *Comput. Phys. Commun.*, 10:343–367, 1975.
- [7] J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Phil. Trans. Royal Soc. London, Series A*, 236, 1937.
- [8] G. Feldman. Multiple measurements and parameters in the unified approach. Technical report, 2000. Talk at the FermiLab Workshop on Confidence Limits.
- [9] K. Cranmer. Statistical challenges for searches for new physics at the LHC, pp. 112–123, 2005.
- [10] C. Chuang and T. L. Lai. Hybrid resampling methods for confidence intervals. *Statist. Sinica*, 10:1–50, 2000. <http://www3.stat.sinica.edu.tw/statistica/oldpdf/A10n11.pdf>.
- [11] M. Walker, B. Sen and M. Woodroffe. On the unified method with nuisance parameters. *Statist. Sinica*, 19:301–314., 2009. <http://www3.stat.sinica.edu.tw/statistica/oldpdf/A19n116.pdf>.
- [12] G. J. Feldman and R. D. Cousins. A Unified approach to the classical statistical analysis of small signals. *Phys. Rev.*, D57:3873–3889, 1998.
- [13] K. S. Cranmer. Kernel estimation in high-energy physics. *Comput. Phys. Commun.*, 136:198–207, 2001.
- [14] A. L. Read. Linear interpolation of histograms. *Nucl. Instrum. Meth.*, A425:357–360, 1999.
- [15] R. Cousins. Probability density functions for positive nuisance parameters. [http://www.physics.ucla.edu/~cousins/stats/cousins\\_lognormal\\_prior.pdf](http://www.physics.ucla.edu/~cousins/stats/cousins_lognormal_prior.pdf).
- [16] W. T. Eadie, D. Drijard, F. E. James, M. Roos, and B. Sadoulet. *Statistical methods in experimental physics*,. American Elsevier Pub. Co, 1st edition.
- [17] R. J. Barlow and C. Beeston. Fitting using finite Monte Carlo samples. *Comput.Phys.Commun.*, 77:219–228, 1993.
- [18] ATLAS Collaboration. Search for the standard model higgs boson in the diphoton decay channel with 4.9 fb<sup>-1</sup> of atlas data at sqrt(s)=7tev. (ATLAS-CONF-2011-161), Dec 2011.
- [19] G. Aad et al. Search for the Standard Model Higgs boson in the diphoton decay channel with 4.9 fb<sup>-1</sup> of pp collisions at sqrt(s)=7 TeV with ATLAS. *Phys.Rev.Lett.*, 108:111803, 2012.
- [20] G. Aad et al. Search for supersymmetry in final states with jets, missing transverse momentum and one isolated lepton in sqrts = 7 TeV pp collisions using 1 fb<sup>-1</sup> of ATLAS data. *Phys.Rev.*, D85:012006, 2012.
- [21] G. Aad et al. Expected Performance of the ATLAS Experiment - Detector, Trigger and Physics. 2009.
- [22] A. Hocker, J. Stelzer, F. Tegenfeldt, H. Voss, K. Voss, et al. TMVA - Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007. TMVA Users Guide: 74 pages, 13 Figures, many code examples

- Report-no: CERN-OPEN-2007-007 Subj-class: Data Analysis, Statistics and Probability.
- [23] G. Punzi. Comments on likelihood fits with variable resolution. *eConf*, C030908:WELT002, 2003.
  - [24] K. Cranmer. in *Proceedings of the PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics*, Geneva, Switzerland, 27 - 29 June, edited by L. Lyons and H. Prosper, CERN-2008-001 (CERN, Geneva, 2008), pp. 47–60, <http://dx.doi.org/10.5170/CERN-2008-001.47>.
  - [25] A. L. Read. Modified frequentist analysis of search results (the CLs method). 2000.
  - [26] A. L. Read. Presentation of search results: the CLs technique. *J. Phys. G: Nucl. Part. Phys.*, 28, 2002.
  - [27] CLs upper limits. [http://en.wikipedia.org/wiki/CLs\\_upper\\_limits](http://en.wikipedia.org/wiki/CLs_upper_limits).
  - [28] G. Cowan, K. Cranmer, E. Gross, and O. Vitells. Power-Constrained Limits. *ArXiv e-prints*, 2011.
  - [29] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9:60–2, 1938.
  - [30] A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54, No. 3:426–482, 1943.
  - [31] M. Woodroofe. <http://people.stat.sfu.ca/lockhart/richard/banff2010/woodroofe.pdf>, 2010.
  - [32] E. Gross and O. Vitells. Trial factors for the look elsewhere effect in high energy physics. *The European Physical Journal C - Particles and Fields*, 70:525–530, 2010. 10.1140/epjc/s10052-010-1470-8.
  - [33] Procedure for the lhc higgs boson search combination in summer 2011. Technical Report ATL-PHYS-PUB-2011-011, CERN, Geneva, 2011.
  - [34] M. Mandelkern. Setting confidence intervals for bounded parameters. *Statistical Science*, 17(2):pp. 149–159, 2002.
  - [35] R. D. Cousins. Negatively Biased Relevant Subsets Induced by the Most-Powerful One-Sided Upper Confidence Limits for a Bounded Physical Parameter. *ArXiv e-prints*, September 2011.
  - [36] A. Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):pp. 269–306, 1962.
  - [37] G. D’Agostini. Bayesian inference in processing experimental data: principles and basic applications. *Reports on Progress in Physics*, 66(9):1383, 2003.
  - [38] R. D. Cousins. Why isn’t every physicist a Bayesian? *Am. J. Phys.*, 63:398, 1995.
  - [39] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.
  - [40] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):pp. 453–461, 1946.
  - [41] L. Demortier, S. Jain, and H. B. Prosper. Reference priors for high energy physics. *Phys. Rev.*, D82:034002, 2010.
  - [42] R. D. Cousins and V. L. Highland. Incorporating systematic uncertainties into an upper limit. *Nucl. Instrum. Meth.*, A320:331–335, 1992. Revised version.
  - [43] J. Conrad and F. Tegenfeldt. Likelihood ratio intervals with Bayesian treatment of uncertainties: Coverage, power and combined experiments, , pp. 93–96, 2005.
  - [44] F. Tegenfeldt and J. Conrad. On Bayesian treatment of systematic uncertainties in confidence interval calculations. *Nucl. Instrum. Meth.*, A539:407–413, 2005.
  - [45] J. Conrad, O. Botner, A. Hallgren, and C. P. de los Heros. Coverage of confidence intervals for Poisson statistics in presence of systematic uncertainties, pp. 58–63, 2002.
  - [46] J. Conrad, O. Botner, A. Hallgren, and C. P. de los Heros. Including systematic uncertainties in confidence interval construction for Poisson statistics. *Phys.Rev.*, D67:012002, 2003.
  - [47] W. A. Rolke, A. M. Lopez, and J. Conrad. Limits and confidence intervals in the presence of

- nuisance parameters. *Nucl. Instrum. Meth.*, A551:493–503, 2005.
- [48] G. C. Hill. Comment on “including systematic uncertainties in confidence interval construction for poisson statistics”. *Phys. Rev. D*, 67:118101, Jun 2003.
- [49] L. Demortier. P values and nuisance parameters. pages 23–33, 2007.
- [50] A. Caldwell, D. Kollar, and K. Kroeninger. Bayesian analysis toolkit. *Comput. Phys. Commun.*, 180, 2009.
- [51] J. Skilling. Nested sampling. *AIP Conference Proceedings*, 735(1):395–405, 2004.
- [52] F. Feroz, M. P. Hobson, and M. Bridges. MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Mon. Not. Roy. Astron. Soc.*, 398:1601–1614, 2009.
- [53] J. O. Berger and J. M. Bernardo. *Bayesian Statistics 4*. Oxford University Press, 1992.
- [54] J. O. Berger and J. M. Bernardo. *Biometrika*, 79:25, 1992.
- [55] J. O. Berger and J. M. Bernardo. *Journal of the American Statistical Association*, 84:200, 1989.
- [56] J. M. Bernardo. *J. R. Statist. Soc. B*, 41:113, 1979.
- [57] D. Casadei. Reference analysis of the signal + background model in counting experiments. *JINST*, 7:P01012, 2012.
- [58] *Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*, Geneva, Switzerland, 17–20 January 2011, edited by H. B. Prosper and L. Lyons, CERN-2011-006 (CERN, Geneva, 2011), <http://dx.doi.org/10.5170/CERN-2011-006>.
- [59] G. D’Agostini. A multidimensional unfolding method based on bayes’ theorem. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 362(2-3):487 – 498, 1995.
- [60] T. Adye. Unfolding algorithms and tests using RooUnfold. 2011.
- [61] B. Malaescu. An Iterative, Dynamically Stabilized(IDS) Method of Data Unfolding. [arxiv:1106.3107], 2011.
- [62] V. Blobel. An Unfolding method for high-energy physics experiments. [hep-ex/0208022], 2002.
- [63] A. Hocker and V. Kartvelishvili. SVD approach to data unfolding. *Nucl. Instrum. Meth.*, A372:469–481, 1996.
- [64] G. Choudalakis. Fully Bayesian Unfolding. page 24, January 2012.
- [65] A. N. Tikhonov. On the solution of improperly posed problems and the method of regularization. *Sov. Math.*, 5:1035, 1963.



## Cosmology

*J. García-Bellido*

Instituto de Física Teórica IFT-UAM/CSIC, Cantoblanco 28049 Madrid, Spain

### Abstract

In these lectures I review the present status of the so-called Standard Cosmological Model, based on the hot Big Bang Theory and the Inflationary Paradigm. I will make special emphasis on the recent developments in observational cosmology, mainly the acceleration of the universe, the precise measurements of the microwave background anisotropies, and the formation of structure like galaxies and clusters of galaxies from tiny primordial fluctuations generated during inflation.

### 1 Introduction

The last ten years have seen the coming of age of Modern Cosmology, a mature branch of science based on the hot Big Bang theory and the Inflationary Paradigm. In particular, we can now define rather precisely a Standard Model of Cosmology, where the basic parameters are determined within small uncertainties, of just a few percent, thanks to a host of experiments and observations. This precision era of cosmology has become possible thanks to important experimental developments in all fronts, from measurements of supernovae at high redshifts to the microwave background anisotropies, as well as to the distribution of matter in galaxies and clusters of galaxies.

In these lecture notes I will first introduce the basic concepts and equations associated with hot Big Bang cosmology, defining the main cosmological parameters and their corresponding relationships. Then I will address in detail the three fundamental observations that have shaped our present knowledge: the recent acceleration of the universe, the distribution of matter on large scales and the anisotropies in the microwave background. Together these observations allow the precise determination of a handful of cosmological parameters, in the context of the inflationary plus cold dark matter paradigm.

### 2 Big Bang Cosmology

Our present understanding of the universe is based upon the successful hot Big Bang theory, which explains its evolution from the first fraction of a second to our present age, around 13.6 billion years later. This theory rests upon four robust pillars, a theoretical framework based on general relativity, as put forward by Albert Einstein [1] and Alexander A. Friedmann [2] in the 1920s, and three basic observational facts: First, the expansion of the universe, discovered by Edwin P. Hubble [3] in the 1930s, as a recession of galaxies at a speed proportional to their distance from us. Second, the relative abundance of light elements, explained by George Gamow [4] in the 1940s, mainly that of helium, deuterium and lithium, which were cooked from the nuclear reactions that took place at around a second to a few minutes after the Big Bang, when the universe was a few times hotter than the core of the sun. Third, the cosmic microwave background (CMB), the afterglow of the Big Bang, discovered in 1965 by Arno A. Penzias and Robert W. Wilson [5] as a very isotropic blackbody radiation at a temperature of about 3 degrees Kelvin, emitted when the universe was cold enough to form neutral atoms, and photons decoupled from matter, approximately 380,000 years after the Big Bang. Today, these observations are confirmed to within a few percent accuracy, and have helped establish the hot Big Bang as the preferred model of the universe.

Modern Cosmology begun as a quantitative science with the advent of Einstein's general relativity and the realization that the geometry of space-time, and thus the general attraction of matter, is

determined by the energy content of the universe [6]

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}. \quad (1)$$

These non-linear equations are simply too difficult to solve without invoking some symmetries of the problem at hand: the universe itself.

We live on Earth, just 8 light-minutes away from our star, the Sun, which is orbiting at 8.5 kpc from the center of our galaxy,<sup>1</sup> the Milky Way, an ordinary galaxy within the Virgo cluster, of size a few Mpc, itself part of a supercluster of size a few 100 Mpc, within the visible universe, approximately 10,000 Mpc in size. Although at small scales the universe looks very inhomogeneous and anisotropic, the deepest galaxy catalogs like 2dF GRS and SDSS suggest that the universe on large scales (beyond the supercluster scales) is very homogeneous and isotropic. Moreover, the cosmic microwave background, which contains information about the early universe, indicates that the deviations from homogeneity and isotropy were just a few parts per million at the time of photon decoupling. Therefore, we can safely impose those symmetries to the universe at large and determine the corresponding evolution equations. The most general metric satisfying homogeneity and isotropy is the Friedmann-Robertson-Walker (FRW) metric, written here in terms of the invariant geodesic distance  $ds^2 = g_{\mu\nu}dx^\mu dx^\nu$  in four dimensions [6]

$$ds^2 = -dt^2 + a^2(t) \left[ \frac{dr^2}{1 - Kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2) \right], \quad (2)$$

characterized by just two quantities, a *scale factor*  $a(t)$ , which determines the physical size of the universe, and a constant  $K$ , which characterizes the *spatial* curvature of the universe,

$${}^{(3)}R = \frac{6K}{a^2(t)} \quad \begin{cases} K = -1 & \text{OPEN} \\ K = 0 & \text{FLAT} \\ K = +1 & \text{CLOSED} \end{cases} \quad (3)$$

Spatially open, flat and closed universes have different three-geometries. Light geodesics on these universes behave differently, and thus could in principle be distinguished observationally, as we shall discuss later. Apart from the three-dimensional spatial curvature, we can also compute a four-dimensional *space-time* curvature,

$${}^{(4)}R = 6\frac{\ddot{a}}{a} + 6\left(\frac{\dot{a}}{a}\right)^2 + 6\frac{K}{a^2}. \quad (4)$$

Depending on the dynamics (and thus on the matter/energy content) of the universe, we will have different possible outcomes of its evolution. The universe may expand for ever, recollapse in the future or approach an asymptotic state in between.

## 2.1 The matter and energy content of the universe

The most general matter fluid consistent with the assumption of homogeneity and isotropy is a perfect fluid, one in which an observer *comoving with the fluid* would see the universe around it as isotropic. The energy momentum tensor associated with such a fluid can be written as [6]

$$T^{\mu\nu} = p g^{\mu\nu} + (p + \rho) U^\mu U^\nu, \quad (5)$$

where  $p(t)$  and  $\rho(t)$  are the pressure and energy density of the fluid at a given time in the expansion, as measured by this comoving observer, and  $U^\mu$  is the comoving four-velocity, satisfying  $U^\mu U_\mu = -1$ . For such a comoving observer, the matter content looks isotropic (in its rest frame),

$$T^\mu{}_\nu = \text{diag}(-\rho(t), p(t), p(t), p(t)). \quad (6)$$

<sup>1</sup>One parallax second (1 pc), *parsec* for short, corresponds to a distance of about 3.26 light-years or  $3.09 \times 10^{18}$  cm.

<sup>2</sup>I am using  $c = 1$  everywhere, unless specified, and a metric signature  $(-, +, +, +)$ .



The conservation of energy ( $T^{\mu\nu}_{;\nu} = 0$ ), a direct consequence of the general covariance of the theory ( $G^{\mu\nu}_{;\nu} = 0$ ), can be written in terms of the FRW metric and the perfect fluid tensor (5) as

$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + p) = 0. \quad (7)$$

In order to find explicit solutions, one has to supplement the conservation equation with an *equation of state* relating the pressure and the density of the fluid,  $p = p(\rho)$ . The most relevant fluids in cosmology are barotropic, i.e. fluids whose pressure is linearly proportional to the density,  $p = w\rho$ , and therefore the speed of sound is constant in those fluids.

We will restrict ourselves in these lectures to three main types of barotropic fluids:

- *Radiation*, with equation of state  $p_R = \rho_R/3$ , associated with relativistic degrees of freedom (i.e. particles with temperatures much greater than their mass). In this case, the energy density of radiation decays as  $\rho_R \sim a^{-4}$  with the expansion of the universe.
- *Matter*, with equation of state  $p_M \simeq 0$ , associated with nonrelativistic degrees of freedom (i.e. particles with temperatures much smaller than their mass). In this case, the energy density of matter decays as  $\rho_M \sim a^{-3}$  with the expansion of the universe.
- *Vacuum energy*, with equation of state  $p_V = -\rho_V$ , associated with quantum vacuum fluctuations. In this case, the vacuum energy density remains constant with the expansion of the universe.

This is all we need in order to solve the Einstein equations. Let us now write the equations of motion of observers comoving with such a fluid in an expanding universe. According to general relativity, these equations can be deduced from the Einstein equations (1), by substituting the FRW metric (2) and the perfect fluid tensor (5). The  $\mu = i$ ,  $\nu = j$  component of the Einstein equations, together with the  $\mu = 0$ ,  $\nu = 0$  component constitute the so-called Friedmann equations,

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho + \frac{\Lambda}{3} - \frac{K}{a^2}, \quad (8)$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p) + \frac{\Lambda}{3}. \quad (9)$$

These equations contain all the relevant dynamics, since the energy conservation equation (7) can be obtained from these.

## 2.2 The Cosmological Parameters

I will now define the most important cosmological parameters. Perhaps the best known is the *Hubble parameter* or rate of expansion today,  $H_0 = \dot{a}/a(t_0)$ . We can write the Hubble parameter in units of  $100 \text{ km s}^{-1}\text{Mpc}^{-1}$ , which can be used to estimate the order of magnitude for the present size and age of the universe,

$$H_0 \equiv 100 h \text{ km s}^{-1}\text{Mpc}^{-1}, \quad (10)$$

$$c H_0^{-1} = 3000 h^{-1} \text{ Mpc}, \quad (11)$$

$$H_0^{-1} = 9.773 h^{-1} \text{ Gyr}. \quad (12)$$

The parameter  $h$  was measured to be in the range  $0.4 < h < 1$  for decades, and only in the last few years has it been found to lie within 4% of  $h = 0.70$ . I will discuss those recent measurements in the next Section.

Using the present rate of expansion, one can define a *critical density*  $\rho_c$ , that which corresponds to a flat universe,

$$\rho_c \equiv \frac{3H_0^2}{8\pi G} = 1.88 h^2 10^{-29} \text{ g/cm}^3 \quad (13)$$

$$= 2.77 h^{-1} 10^{11} M_{\odot}/(h^{-1} \text{Mpc})^3 \quad (14)$$

$$= 11.26 h^2 \text{ protons}/\text{m}^3, \quad (15)$$

where  $M_{\odot} = 1.989 \times 10^{33} \text{ g}$  is a solar mass unit. The critical density  $\rho_c$  corresponds to approximately 6 protons per cubic meter, certainly a very dilute fluid!

In terms of the critical density it is possible to define the density parameter

$$\Omega_0 \equiv \frac{8\pi G}{3H_0^2} \rho(t_0) = \frac{\rho}{\rho_c}(t_0), \quad (16)$$

whose sign can be used to determine the spatial (three-)curvature. Closed universes ( $K = +1$ ) have  $\Omega_0 > 1$ , flat universes ( $K = 0$ ) have  $\Omega_0 = 1$ , and open universes ( $K = -1$ ) have  $\Omega_0 < 1$ , no matter what are the individual components that sum up to the density parameter.

In particular, we can define the individual ratios  $\Omega_i \equiv \rho_i/\rho_c$ , for matter, radiation, cosmological constant and even curvature, today,

$$\Omega_M = \frac{8\pi G \rho_M}{3H_0^2} \quad \Omega_R = \frac{8\pi G \rho_R}{3H_0^2} \quad (17)$$

$$\Omega_{\Lambda} = \frac{\Lambda}{3H_0^2} \quad \Omega_K = -\frac{K}{a_0^2 H_0^2}. \quad (18)$$

For instance, we can evaluate today the radiation component  $\Omega_R$ , corresponding to relativistic particles, from the density of microwave background photons,  $\rho_{\text{CMB}} = \pi^2 k^4 T_{\text{CMB}}^4 / (15 \hbar^3 c^3) = 4.5 \times 10^{-34} \text{ g}/\text{cm}^3$ , which gives  $\Omega_{\text{CMB}} = 2.4 \times 10^{-5} h^{-2}$ . Three approximately massless neutrinos would contribute a similar amount. Therefore, we can safely neglect the contribution of relativistic particles to the total density of the universe today, which is dominated either by non-relativistic particles (baryons, dark matter or massive neutrinos) or by a cosmological constant, and write the rate of expansion in terms of its value today, as

$$H^2(a) = H_0^2 \left( \Omega_R \frac{a_0^4}{a^4} + \Omega_M \frac{a_0^3}{a^3} + \Omega_{\Lambda} + \Omega_K \frac{a_0^2}{a^2} \right). \quad (19)$$

An interesting consequence of these definitions is that one can now write the Friedmann equation today,  $a = a_0$ , as a *cosmic sum rule*,

$$1 = \Omega_M + \Omega_{\Lambda} + \Omega_K, \quad (20)$$

where we have neglected  $\Omega_R$  today. That is, in the context of a FRW universe, the total fraction of matter density, cosmological constant and spatial curvature today must add up to one. For instance, if we measure one of the three components, say the spatial curvature, we can deduce the sum of the other two.

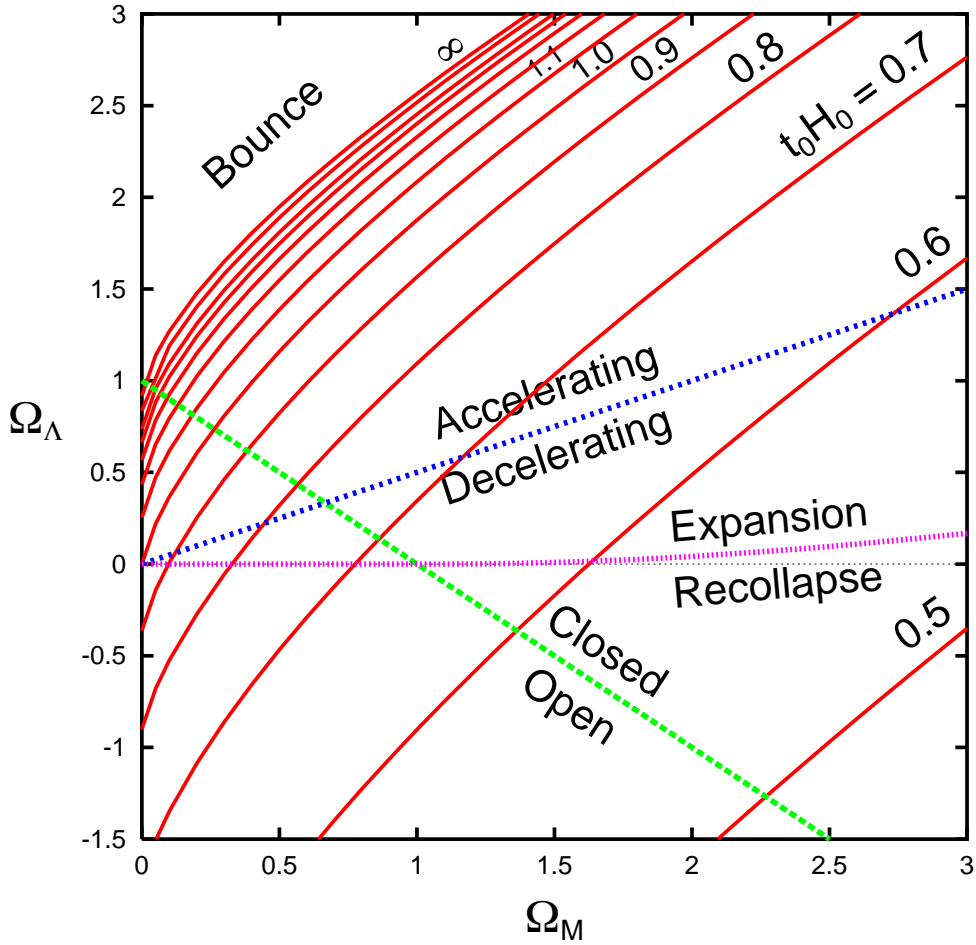
Looking now at the second Friedmann equation (9), we can define another basic parameter, the *deceleration parameter*,

$$q_0 = -\frac{a \ddot{a}}{\dot{a}^2}(t_0) = \frac{4\pi G}{3H_0^2} \left[ \rho(t_0) + 3p(t_0) \right], \quad (21)$$

defined so that it is positive for ordinary matter and radiation, expressing the fact that the universe expansion should slow down due to the gravitational attraction of matter. We can write this parameter using the definitions of the density parameter for known and unknown fluids (with density  $\Omega_x$  and arbitrary equation of state  $w(z)$ ) as

$$q_0 = \Omega_R + \frac{1}{2}\Omega_M - \Omega_{\Lambda} + \frac{1}{2} \sum_x (1 + 3w) \Omega_x. \quad (22)$$

Uniform expansion corresponds to  $q_0 = 0$  and requires a cancellation between the matter and vacuum energies. For matter domination,  $q_0 > 0$ , while for vacuum domination,  $q_0 < 0$ . As we will see in a



**Fig. 1:** Parameter space  $(\Omega_M, \Omega_\Lambda)$ . The green (dashed) line  $\Omega_\Lambda = 1 - \Omega_M$  corresponds to a flat universe,  $\Omega_K = 0$ , separating open from closed universes. The blue (dotted) line  $\Omega_\Lambda = \Omega_M/2$  corresponds to uniform expansion,  $q_0 = 0$ , separating accelerating from decelerating universes. The violet (dot-dashed) line corresponds to critical universes, separating eternal expansion from recollapse in the future. Finally, the red (continuous) lines correspond to  $t_0 H_0 = 0.5, 0.6, \dots, \infty$ , beyond which the universe has a bounce.

moment, we are at present probing the time dependence of the deceleration parameter and can determine with some accuracy the moment at which the universe went from a decelerating phase, dominated by dark matter, into an acceleration phase at present, which seems to indicate the dominance of some kind of vacuum energy.

### 2.3 The $(\Omega_M, \Omega_\Lambda)$ plane

Now that we know that the universe is accelerating, one can parametrize the matter/energy content of the universe with just two components: the matter, characterized by  $\Omega_M$ , and the vacuum energy  $\Omega_\Lambda$ . Different values of these two parameters completely specify the universe evolution. It is thus natural to plot the results of observations in the plane  $(\Omega_M, \Omega_\Lambda)$ , in order to check whether we arrive at a consistent picture of the present universe from several different angles (different sets of cosmological observations).

Moreover, different regions of this plane specify different behaviors of the universe. The boundaries between regions are well defined curves that can be computed for a given model. I will now describe the various regions and boundaries.

- *Uniform expansion* ( $q_0 = 0$ ). Corresponds to the line  $\Omega_\Lambda = \Omega_M/2$ . Points above this line correspond to universes that are accelerating today, while those below correspond to decelerating universes, in particular the old cosmological model of Einstein-de Sitter (EdS), with  $\Omega_\Lambda = 0$ ,  $\Omega_M = 1$ . Since 1998, all the data from Supernovae of type Ia appear above this line, many standard deviations away from EdS universes.
- *Flat universe* ( $\Omega_K = 0$ ). Corresponds to the line  $\Omega_\Lambda = 1 - \Omega_M$ . Points to the right of this line correspond to closed universes, while those to the left correspond to open ones. In the last few years we have mounting evidence that the universe is spatially flat (in fact Euclidean).
- *Bounce* ( $t_0 H_0 = \infty$ ). Corresponds to a complicated function of  $\Omega_\Lambda(\Omega_M)$ , normally expressed as an integral equation, where

$$t_0 H_0 = \int_0^1 da [1 + \Omega_M(1/a - 1) + \Omega_\Lambda(a^2 - 1)]^{-1/2}$$

is the product of the age of the universe and the present rate of expansion. Points above this line correspond to universes that have contracted in the past and have later rebounded. At present, these universes are ruled out by observations of galaxies and quasars at high redshift (up to  $z = 10$ ).

- *Critical Universe* ( $H = \dot{H} = 0$ ,  $\ddot{H} > 0$ ). Corresponds to the boundary between eternal expansion in the future and recollapse. For  $\Omega_M \leq 1$ , it is simply the line  $\Omega_\Lambda = 0$ , but for  $\Omega_M > 1$ , it is a more complicated curve,

$$\Omega_\Lambda = 4\Omega_M \sin^3 \left[ \frac{1}{3} \arcsin \left( \frac{\Omega_M - 1}{\Omega_M} \right) \right] \simeq \frac{4}{27} \frac{(\Omega_M - 1)^3}{\Omega_M^2}.$$

These critical solutions are asymptotic to the EdS model.

These boundaries, and the regions they delimit, can be seen in Fig. 1, together with the lines of equal  $t_0 H_0$  values.

In summary, the basic cosmological parameters that are now being hunted by a host of cosmological observations are the following: the present rate of expansion  $H_0$ ; the age of the universe  $t_0$ ; the deceleration parameter  $q_0$ ; the spatial curvature  $\Omega_K$ ; the matter content  $\Omega_M$ ; the vacuum energy  $\Omega_\Lambda$ ; the baryon density  $\Omega_B$ ; the neutrino density  $\Omega_\nu$ , and many other that characterize the perturbations responsible for the large scale structure (LSS) and the CMB anisotropies.

## 2.4 The accelerating universe

Let us first describe the effect that the expansion of the universe has on the objects that live in it. In the absence of other forces but those of gravity, the trajectory of a particle is given by general relativity in terms of the geodesic equation

$$\frac{du^\mu}{ds} + \Gamma_{\nu\lambda}^\mu u^\nu u^\lambda = 0, \quad (23)$$

where  $u^\mu = (\gamma, \gamma v^i)$ , with  $\gamma^2 = 1 - v^2$  and  $v^i$  is the peculiar velocity. Here  $\Gamma_{\nu\lambda}^\mu$  is the Christoffel connection [6], whose only non-zero component is  $\Gamma_{ij}^0 = (\dot{a}/a) g_{ij}$ ; substituting into the geodesic equation, we obtain  $|\vec{v}| \propto 1/a$ , and thus the particle's momentum decays with the expansion like  $p \propto 1/a$ . In the case of a photon, satisfying the de Broglie relation  $p = h/\lambda$ , one obtains the well known *photon redshift*

$$\frac{\lambda_1}{\lambda_0} = \frac{a(t_1)}{a(t_0)} \Rightarrow z \equiv \frac{\lambda_0 - \lambda_1}{\lambda_1} = \frac{a_0}{a_1} - 1, \quad (24)$$

where  $\lambda_0$  is the wavelength measured by an observer at time  $t_0$ , while  $\lambda_1$  is the wavelength emitted when the universe was younger ( $t_1 < t_0$ ). Normally we measure light from stars in distant galaxies and compare their observed spectra with our laboratory (restframe) spectra. The fraction (24) then gives the

redshift  $z$  of the object. We are assuming, of course, that both the emitted and the restframe spectra are identical, so that we can actually measure the effect of the intervening expansion, i.e. the growth of the scale factor from  $t_1$  to  $t_0$ , when we compare the two spectra. Note that if the emitting galaxy and our own participated in the expansion, i.e. if our measuring rods (our rulers) also expanded with the universe, we would see no effect! The reason we can measure the redshift of light from a distant galaxy is because our galaxy is a gravitationally bounded object that has decoupled from the expansion of the universe. It is the distance between galaxies that changes with time, not the sizes of galaxies, nor the local measuring rods.

We can now evaluate the relationship between physical distance and redshift as a function of the rate of expansion of the universe. Because of homogeneity we can always choose our position to be at the origin  $r = 0$  of our spatial section. Imagine an object (a star) emitting light at time  $t_1$ , at coordinate distance  $r_1$  from the origin. Because of isotropy we can ignore the angular coordinates  $(\theta, \phi)$ . Then the physical distance, to first order, will be  $d = a_0 r_1$ . Since light travels along null geodesics [6], we can write  $0 = -dt^2 + a^2(t) dr^2/(1 - Kr^2)$ , and therefore,

$$\int_{t_1}^{t_0} \frac{dt}{a(t)} = \int_0^{r_1} \frac{dr}{\sqrt{1 - Kr^2}} \equiv f(r_1) = \begin{cases} \arcsin r_1 & K = 1 \\ r_1 & K = 0 \\ \operatorname{arcsinh} r_1 & K = -1 \end{cases} \quad (25)$$

If we now Taylor expand the scale factor to first order,

$$\frac{1}{1+z} = \frac{a(t)}{a_0} = 1 + H_0(t - t_0) + \mathcal{O}(t - t_0)^2, \quad (26)$$

we find, to first approximation,

$$r_1 \approx f(r_1) = \frac{1}{a_0}(t_0 - t_1) + \dots = \frac{z}{a_0 H_0} + \dots$$

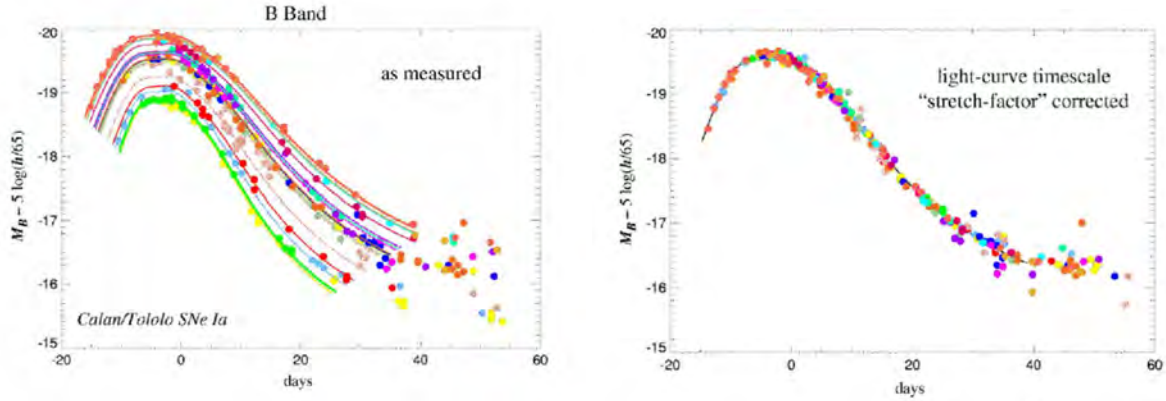
Putting all together we find the famous Hubble law

$$H_0 d = a_0 H_0 r_1 = z \simeq vc, \quad (27)$$

which is just a kinematical effect (we have not included yet any dynamics, i.e. the matter content of the universe). Note that at low redshift ( $z \ll 1$ ), one is tempted to associate the observed change in wavelength with a Doppler effect due to a hypothetical recession velocity of the distant galaxy. This is only an approximation. In fact, the redshift cannot be ascribed to the relative velocity of the distant galaxy because in general relativity (i.e. in curved spacetimes) one cannot compare velocities through parallel transport, since the value depends on the path! If the distance to the galaxy is small, i.e.  $z \ll 1$ , the physical spacetime is not very different from Minkowsky and such a comparison is approximately valid. As  $z$  becomes of order one, such a relation is manifestly false: galaxies cannot travel at speeds greater than the speed of light; it is the stretching of spacetime which is responsible for the observed redshift.

Hubble's law has been confirmed by observations ever since the 1920s, with increasing precision, which have allowed cosmologists to determine the Hubble parameter  $H_0$  with less and less systematic errors. Nowadays, the best determination of the Hubble parameter was made by the Hubble Space Telescope Key Project [8],  $H_0 = 72 \pm 8$  km/s/Mpc. This determination is based on objects at distances up to 500 Mpc, corresponding to redshifts  $z \leq 0.1$ .

Nowadays, we are beginning to probe much greater distances, corresponding to  $z \simeq 1$ , thanks to type Ia supernovae. These are white dwarf stars at the end of their life cycle that accrete matter from a companion until they become unstable and violently explode in a natural thermonuclear explosion that out-shines their progenitor galaxy. The intensity of the distant flash varies in time, it takes about



**Fig. 2:** The Type Ia supernovae observed nearby show a relationship between their absolute luminosity and the timescale of their light curve: the brighter supernovae are slower and the fainter ones are faster. A simple linear relation between the absolute magnitude and a “stretch factor” multiplying the light curve timescale fits the data quite well. From Ref. [7].

three weeks to reach its maximum brightness and then it declines over a period of months. Although the maximum luminosity varies from one supernova to another, depending on their original mass, their environment, etc., there is a pattern: brighter explosions last longer than fainter ones. By studying the characteristic light curves, see Fig. 2, of a reasonably large statistical sample, cosmologists from the Supernova Cosmology Project [7] and the High-redshift Supernova Project [9], are now quite confident that they can use this type of supernova as a standard candle. Since the light coming from some of these rare explosions has travelled a large fraction of the size of the universe, one expects to be able to infer from their distribution the spatial curvature and the rate of expansion of the universe.

The connection between observations of high redshift supernovae and cosmological parameters is done via the luminosity distance, defined as the distance  $d_L$  at which a source of absolute luminosity (energy emitted per unit time)  $\mathcal{L}$  gives a flux (measured energy per unit time and unit area of the detector)  $\mathcal{F} = \mathcal{L}/4\pi d_L^2$ . One can then evaluate, within a given cosmological model, the expression for  $d_L$  as a function of redshift [10],

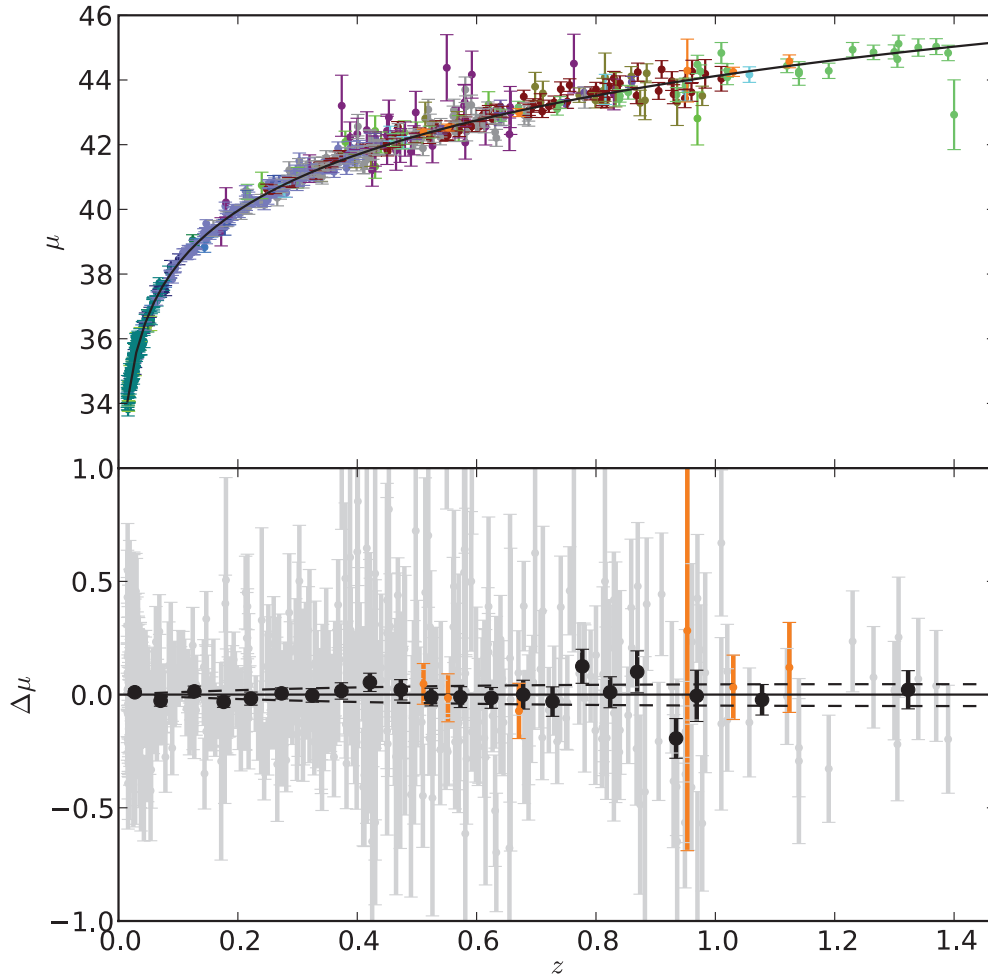
$$H_0 d_L(z) = \frac{(1+z)}{|\Omega_K|^{1/2}} \text{sinn} \left[ \int_0^z \frac{|\Omega_K|^{1/2} dz'}{\sqrt{(1+z')^2(1+z'\Omega_M) - z'(2+z')\Omega_\Lambda}} \right], \quad (28)$$

where  $\text{sinn}(x) = x$  if  $K = 0$ ;  $\sin(x)$  if  $K = +1$  and  $\sinh(x)$  if  $K = -1$ , and we have used the cosmic sum rule (20).

Astronomers measure the relative luminosity of a distant object in terms of what they call the effective magnitude, which has a peculiar relation with distance,

$$m(z) \equiv M + 5 \log_{10} \left[ \frac{d_L(z)}{\text{Mpc}} \right] + 25 = \bar{M} + 5 \log_{10}[H_0 d_L(z)]. \quad (29)$$

Since 1998, several groups have obtained serious evidence that high redshift supernovae appear fainter than expected for either an open ( $\Omega_M < 1$ ) or a flat ( $\Omega_M = 1$ ) universe, see Fig. 3. In fact, the universe appears to be accelerating instead of decelerating, as was expected from the general attraction of matter, see Eq. (22); something seems to be acting as a repulsive force on very large scales. The most natural explanation for this is the presence of a cosmological constant, a diffuse vacuum energy that permeates all space and, as explained above, gives the universe an acceleration that tends to separate gravitationally bound systems from each other. The best-fit results from the Supernova Cosmology Project [11] give a



**Fig. 3:** The Hubble diagram in linear redshift scale. Supernovae with  $\Delta z < 0.05$  of each other have been weighted-averaged binned. The solid curve represents the best-fit flat universe model, ( $\Omega_M = 0.29$ ,  $\Omega_\Lambda = 0.71$ ). Lower panel: Residuals of the averaged data relative to a  $\Lambda$ CDM universe with fiducial cosmology. Dashed lines correspond to  $w = -1 \pm 0.1$ . From Ref. [13].

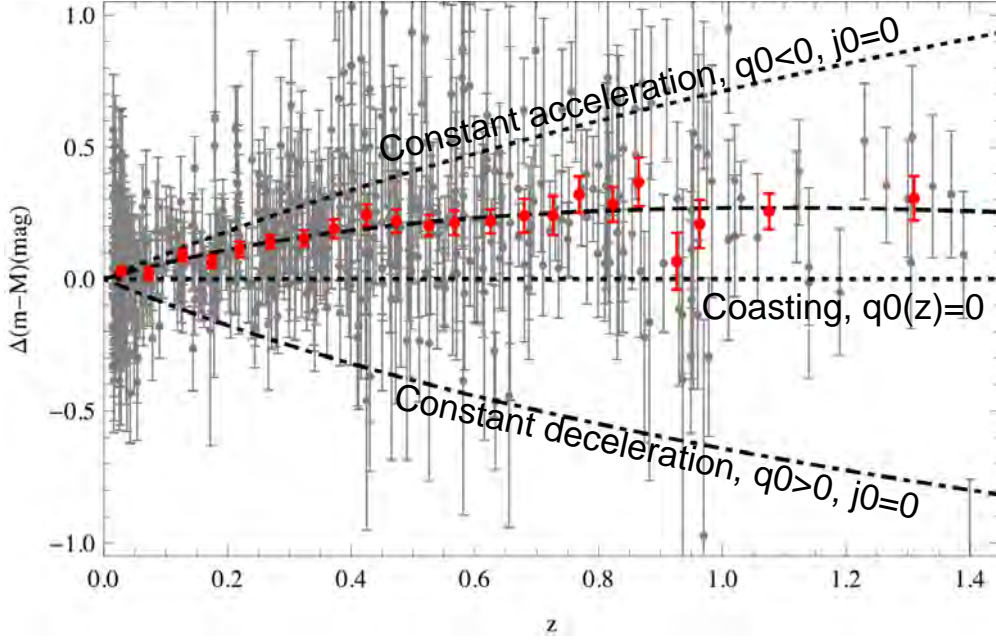
linear combination

$$0.8 \Omega_M - 0.6 \Omega_\Lambda = -0.16 \pm 0.05 \quad (1\sigma),$$

which is now many sigma away from an EdS model with  $\Lambda = 0$ . In particular, for a flat universe this gives

$$\Omega_\Lambda = 0.71 \pm 0.05 \quad \text{and} \quad \Omega_M = 0.29 \pm 0.05 \quad (1\sigma).$$

Surprising as it may seem, arguments for a significant dark energy component of the universe were proposed long before these observations, in order to accommodate the ages of globular clusters, as well as a flat universe with a matter content below critical, which was needed in order to explain the observed distribution of galaxies, clusters and voids.



**Fig. 4:** The Supernovae Ia residual Hubble diagram. Upper panel: Ground-based discoveries are represented by diamonds, HST-discovered SNe Ia are shown as filled circles. Lower panel: The same but with weighted averaged in fixed redshift bins. Kinematic models of the expansion history are shown relative to an eternally coasting model  $q(z) = 0$ . Adapted from Ref. [13].

Taylor expanding the scale factor to third order,

$$\frac{a(t)}{a_0} = 1 + H_0(t - t_0) - \frac{q_0}{2!} H_0^2 (t - t_0)^2 + \frac{j_0}{3!} H_0^3 (t - t_0)^3 + \mathcal{O}(t - t_0)^4, \quad (30)$$

where

$$q_0 = -\frac{\ddot{a}}{aH^2}(t_0) = \frac{1}{2} \sum_i (1 + 3w_i) \Omega_i = \frac{1}{2} \Omega_M - \Omega_\Lambda, \quad (31)$$

$$j_0 = +\frac{\dddot{a}}{aH^3}(t_0) = \frac{1}{2} \sum_i (1 + 3w_i)(2 + 3w_i) \Omega_i = \Omega_M + \Omega_\Lambda, \quad (32)$$

are the deceleration and “jerk” parameters. Substituting into Eq. (28) we find

$$H_0 d_L(z) = z + \frac{1}{2}(1 - q_0) z^2 - \frac{1}{6}(1 - q_0 - 3q_0^2 + j_0) z^3 + \mathcal{O}(z^4). \quad (33)$$

This expression goes beyond the leading linear term, corresponding to the Hubble law, into the second and third order terms, which are sensitive to the cosmological parameters  $\Omega_M$  and  $\Omega_\Lambda$ . It is only recently that cosmological observations have gone far enough back into the early universe that we can begin to probe these terms, see Fig. 4.

This extra component of the critical density would have to resist gravitational collapse, otherwise it would have been detected already as part of the energy in the halos of galaxies. However, if most of the energy of the universe resists gravitational collapse, it is impossible for structure in the universe to grow. This dilemma can be resolved if the hypothetical dark energy was negligible in the past and only recently became the dominant component. According to general relativity, this requires that the dark energy have



negative pressure, since the ratio of dark energy to matter density goes like  $a(t)^{-3p/\rho}$ . This argument would rule out almost all of the usual suspects, such as cold dark matter, neutrinos, radiation, and kinetic energy, since they all have zero or positive pressure. Thus, we expect something like a cosmological constant, with a negative pressure,  $p \approx -\rho$ , to account for the missing energy.

However, if the universe was dominated by dark matter in the past, in order to form structure, and only recently became dominated by dark energy, we must be able to see the effects of the transition from the deceleration into the acceleration phase in the luminosity of distant type Ia supernovae. This has been searched for since 1998, when the first convincing results on the present acceleration appeared. However, only recently [12] do we have clear evidence of this transition point in the evolution of the universe. This *coasting point* is defined as the time, or redshift, at which the deceleration parameter vanishes,

$$q(z) = -1 + \frac{d \ln H(z)}{d \ln(1+z)} = 0, \quad (34)$$

where

$$H(z) = H_0 \left[ \Omega_M (1+z)^3 + \Omega_x e^{3 \int_0^z (1+w(z')) \frac{dz'}{1+z'}} + \Omega_K (1+z)^2 \right]^{1/2}, \quad (35)$$

and we have assumed that the dark energy is parametrized by a density  $\Omega_x$  today, with a redshift-dependent equation of state,  $w(z)$ , not necessarily equal to  $-1$ . Of course, in the case of a true cosmological constant, this reduces to the usual expression.

Let us suppose for a moment that the barotropic parameter  $w$  is constant, then the coasting redshift can be determined from

$$q(z) = \frac{1}{2} \left[ \frac{\Omega_M + (1+3w) \Omega_x (1+z)^{3w}}{\Omega_M + \Omega_x (1+z)^{3w} + \Omega_K (1+z)^{-1}} \right] = 0, \quad (36)$$

$$\Rightarrow z_c = \left( \frac{(3|w|-1)\Omega_x}{\Omega_M} \right)^{\frac{1}{3|w|}} - 1, \quad (37)$$

which, in the case of a true cosmological constant, reduces to

$$z_c = \left( \frac{2\Omega_\Lambda}{\Omega_M} \right)^{1/3} - 1. \quad (38)$$

When substituting  $\Omega_\Lambda \simeq 0.71$  and  $\Omega_M \simeq 0.29$ , one obtains  $z_c \simeq 0.7$ , in excellent agreement with recent observations [13].

Now, if we have to live with this vacuum energy, we might as well try to understand its origin. For the moment it is a complete mystery, perhaps the biggest mystery we have in physics today [14]. We measure its value but we don't understand why it has the value it has. In fact, if we naively predict it using the rules of quantum mechanics, we find a number that is many (many!) orders of magnitude off the mark. Let us describe this calculation in some detail. In non-gravitational physics, the zero-point energy of the system is irrelevant because forces arise from gradients of potential energies. However, we know from general relativity that even a constant energy density gravitates. Let us write down the most general energy momentum tensor compatible with the symmetries of the metric and that is covariantly conserved. This is precisely of the form  $T_{\mu\nu}^{(vac)} = p_V g_{\mu\nu} = -\rho_V g_{\mu\nu}$ , see Fig. 5. Substituting into the Einstein equations (1), we see that the cosmological constant and the vacuum energy are completely equivalent,  $\Lambda = 8\pi G \rho_V$ , so we can measure the vacuum energy with the observations of the acceleration of the universe, which tells us that  $\Omega_\Lambda \simeq 0.7$ .

On the other hand, we can estimate the contribution to the vacuum energy coming from the quantum mechanical zero-point energy of the quantum oscillators associated with the fluctuations of all quantum fields,

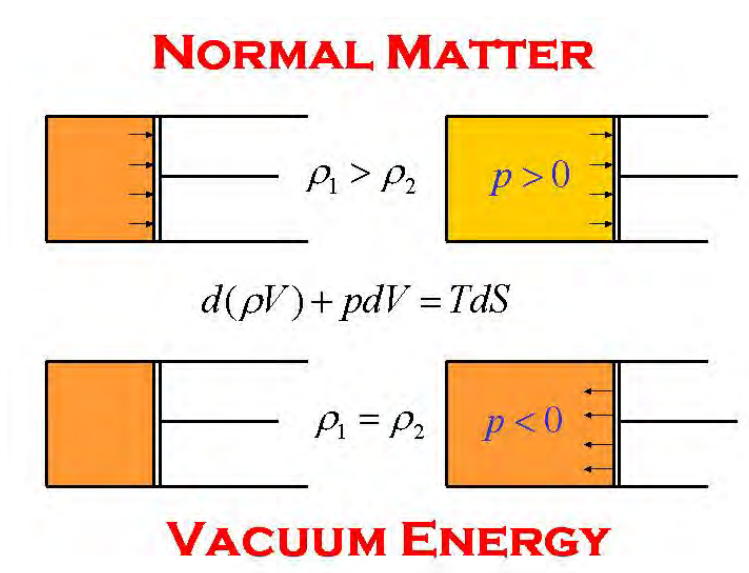
$$\rho_V^{th} = \sum_i \int_0^{\Lambda_{UV}} \frac{d^3 k}{(2\pi)^3} \frac{1}{2} \hbar \omega_i(k) = \frac{\hbar \Lambda_{UV}^4}{16\pi^2} \sum_i (-1)^F N_i + \mathcal{O}(m_i^2 \Lambda_{UV}^2), \quad (39)$$

where  $\Lambda_{UV}$  is the ultraviolet cutoff signaling the scale of new physics. Taking the scale of quantum gravity,  $\Lambda_{UV} = M_{Pl}$ , as the cutoff, and barring any fortuitous cancellations, then the theoretical expectation (39) appears to be 120 orders of magnitude larger than the observed vacuum energy associated with the acceleration of the universe,

$$\rho_V^{th} \simeq 1.4 \times 10^{74} \text{ GeV}^4 = 3.2 \times 10^{91} \text{ g/cm}^3, \tag{40}$$

$$\rho_V^{obs} \simeq 0.7 \rho_c = 0.66 \times 10^{-29} \text{ g/cm}^3 = 2.9 \times 10^{-11} \text{ eV}^4. \tag{41}$$

Even if we assumed that the ultraviolet cutoff associated with quantum gravity was as low as the electroweak scale, the theoretical expectation would still be 60 orders of magnitude too big. This is by far the worst mismatch between theory and observations in all of science. There must be something seriously wrong in our present understanding of gravity at the most fundamental level. Perhaps we don't understand the vacuum and its energy does not gravitate after all, or perhaps we need to impose a new principle (or a symmetry) at the quantum gravity level to accommodate such a flagrant mismatch.



**Fig. 5:** Ordinary matter dilutes as it expands. According to the second law of Thermodynamics, its pressure on the walls should be positive, which exerts a force, and energy is lost in the expansion. On the other hand, vacuum energy is always the same, independent of the volume of the region, and thus, according to the second law, its pressure must be negative and of the same magnitude as the energy density. This negative pressure means that the volume tends to increase more and more rapidly, which explains the exponential expansion of the universe dominated by a cosmological constant.

In the meantime, one can at least parametrize our ignorance by making variations on the idea of a *constant* vacuum energy. Let us assume that it actually evolves slowly with time. In that case, we do not expect the equation of state  $p = -\rho$  to remain true, but instead we expect the barotropic parameter  $w(z)$  to depend on redshift. Such phenomenological models have been proposed, and until recently produced results that were compatible with  $w = -1$  today, but with enough uncertainty to speculate on alternatives to a truly constant vacuum energy. However, with the recent supernovae results [12], there seems to be little space for variations, and models of a time-dependent vacuum energy are less and less favoured. In the near future, the SNAP satellite [15] will measure several thousand supernovae at high redshift and therefore map the redshift dependence of both the dark energy density and its equation of state with great precision. This will allow a much better determination of the cosmological parameters  $\Omega_M$  and  $\Omega_\Lambda$ .

## 2.5 Thermodynamics of an expanding plasma

In this section I will describe the main concepts associated with ensembles of particles in thermal equilibrium and the brief periods in which the universe fell out of equilibrium. To begin with, let me make contact between the covariant energy conservation law (7) and the second law of thermodynamics,

$$T dS = dU + p dV, \quad (42)$$

where  $U = \rho V$  is the total energy of the fluid, and  $p = w \rho$  is its barotropic pressure. Taking a comoving volume for the universe,  $V = a^3$ , we find

$$T \frac{dS}{dt} = \frac{d}{dt}(\rho a^3) + p \frac{d}{dt}(a^3) = 0, \quad (43)$$

where we have used (7). Therefore, entropy is conserved during the expansion of the universe,  $dS = 0$ ; i.e., the expansion is adiabatic even in those epochs in which the equation of state changes, like in the matter-radiation transition (not a proper phase transition). Using (7), we can write

$$\frac{d}{dt} \ln(\rho a^3) = -3H w. \quad (44)$$

Thus, our universe expands like a gaseous fluid in thermal equilibrium at a temperature  $T$ . This temperature decreases like that of any expanding fluid, in a way that is inversely proportional to the cubic root of the volume. This implies that in the past the universe was necessarily denser and hotter. As we go back in time we reach higher and higher temperatures, which implies that the mean energy of plasma particles is larger and thus certain fundamental reactions are now possible and even common, giving rise to processes that today we can only attain in particle physics accelerators. That is the reason why it is so important, for the study of early universe, to know the nature of the fundamental interactions at high energies, and the basic connection between cosmology and high energy particle physics. However, I should clarify a misleading statement that is often used: “high energy particle physics colliders reproduce the early universe” by inducing collisions among relativistic particles. Although the energies of some of the interactions at those collisions reach similar values as those attained in the early universe, the physical conditions are rather different. The interactions within the detectors of the great particle physics accelerators occur typically in the perturbative regime, locally, and very far from thermal equilibrium, lasting a minute fraction of a second; on the other hand, the same interactions occurred within a hot plasma in equilibrium in the early universe while it was expanding adiabatically and its duration could be significantly larger, with a distribution in energy that has nothing to do with those associated with particle accelerators. What is true, of course, is that the fundamental parameters corresponding to those interactions – masses and couplings – are assumed to be the same, and therefore present terrestrial experiments can help us imagine what it could have been like in the early universe, and make predictions about the evolution of the universe, in the context of an expanding plasma at high temperatures and high densities, and in thermal equilibrium.

### 2.5.1 Fluids in thermal equilibrium

In order to understand the thermodynamical behaviour of a plasma of different species of particles at high temperatures we will consider a gas of particles with  $g$  internal degrees of freedom weakly interacting. The degrees of freedom corresponding to the different particles can be seen in Table 1. For example, leptons and quarks have 4 degrees of freedom since they correspond to the two helicities for both particle and antiparticle. However, the nature of neutrinos is still unknown. If they happen to be Majorana fermions, then they would be their own antiparticle and the number of degrees of freedom would reduce to 2. For photons and gravitons (without mass) their 2 d.o.f. correspond to their states of polarization. The 8 gluons (also without mass) are the gauge bosons responsible for the strong interaction between

Particle	Spin	Degrees of freedom ( $g$ )	Nature
Higgs	0	1	Massive scalar
photon	1	2	Massless vector
graviton	2	2	Massless tensor
gluon	1	2	Massless vector
$W$ y $Z$	1	3	Massive vector
leptons & quarks	1/2	4	Dirac Fermion
neutrinos	1/2	4 (2)	Dirac (Majorana) Fermion

**Table 1:** The internal degrees of freedom of various fundamental particles.

quarks, and also have 2 d.o.f. each. The vector bosons  $W^\pm$  and  $Z^0$  are massive and thus, apart from the transverse components of the polarization, they also have longitudinal components.

For each of these particles we can compute the number density  $n$ , the energy density  $\rho$  and the pressure  $p$ , in thermal equilibrium at a given temperature  $T$ ,

$$n = g \int \frac{d^3 \mathbf{p}}{(2\pi)^3} f(\mathbf{p}), \quad (45)$$

$$\rho = g \int \frac{d^3 \mathbf{p}}{(2\pi)^3} E(\mathbf{p}) f(\mathbf{p}), \quad (46)$$

$$p = g \int \frac{d^3 \mathbf{p}}{(2\pi)^3} \frac{|\mathbf{p}|^2}{3E} f(\mathbf{p}), \quad (47)$$

where the energy is given by  $E^2 = |\mathbf{p}|^2 + m^2$  and the momentum distribution in thermal (kinetic) equilibrium is

$$f(\mathbf{p}) = \frac{1}{e^{(E-\mu)/T} \pm 1} \quad \begin{cases} -1 & \text{Bose - Einstein} \\ +1 & \text{Fermi - Dirac} \end{cases} \quad (48)$$

The chemical potential  $\mu$  is conserved in these reactions if they are in thermal equilibrium. For example, for reactions of the type  $i + j \longleftrightarrow k + l$ , we have  $\mu_i + \mu_j = \mu_k + \mu_l$ . For example, the chemical potential of the photon vanishes  $\mu_\gamma = 0$ , and thus particles and antiparticles have opposite chemical potentials.

From the equilibrium distributions one can obtain the number density  $n$ , the energy  $\rho$  and the pressure  $p$ , of a particle of mass  $m$  with chemical potential  $\mu$  at the temperature  $T$ ,

$$n = \frac{g}{2\pi^2} \int_m^\infty dE \frac{E(E^2 - m^2)^{1/2}}{e^{(E-\mu)/T} \pm 1}, \quad (49)$$

$$\rho = \frac{g}{2\pi^2} \int_m^\infty dE \frac{E^2(E^2 - m^2)^{1/2}}{e^{(E-\mu)/T} \pm 1}, \quad (50)$$

$$p = \frac{g}{6\pi^2} \int_m^\infty dE \frac{(E^2 - m^2)^{3/2}}{e^{(E-\mu)/T} \pm 1}. \quad (51)$$

For a non-degenerate ( $\mu \ll T$ ) relativistic gas ( $m \ll T$ ), we find

$$n = \frac{g}{2\pi^2} \int_0^\infty \frac{E^2 dE}{e^{E/T} \pm 1} = \begin{cases} \frac{\zeta(3)}{\pi^2} g T^3 & \text{Bosons} \\ \frac{3}{4} \frac{\zeta(3)}{\pi^2} g T^3 & \text{Fermions} \end{cases}, \quad (52)$$

$$\rho = \frac{g}{2\pi^2} \int_0^\infty \frac{E^3 dE}{e^{E/T} \pm 1} = \begin{cases} \frac{\pi^2}{30} g T^4 & \text{Bosons} \\ \frac{7}{8} \frac{\pi^2}{30} g T^4 & \text{Fermions} \end{cases}, \quad (53)$$

$$p = \frac{1}{3}\rho, \quad (54)$$

where  $\zeta(3) = 1.20206\dots$  is the Riemann Zeta function. For relativistic fluids, the energy density per particle is

$$\langle E \rangle \equiv \frac{\rho}{n} = \begin{cases} \frac{\pi^4}{30\zeta(3)} T \simeq 2.701 T & \text{Bosons} \\ \frac{7\pi^4}{180\zeta(3)} T \simeq 3.151 T & \text{Fermions} \end{cases} \quad (55)$$

For relativistic bosons or fermions with  $\mu < 0$  and  $|\mu| < T$ , we have

$$n = \frac{g}{\pi^2} T^3 e^{\mu/T}, \quad (56)$$

$$\rho = \frac{3g}{\pi^2} T^4 e^{\mu/T}, \quad (57)$$

$$p = \frac{1}{3}\rho. \quad (58)$$

For a bosonic particle, a positive chemical potential,  $\mu > 0$ , indicates the presence of a Bose-Einstein condensate, and should be treated separately from the rest of the modes.

On the other hand, for a non-relativistic gas ( $m \gg T$ ), with arbitrary chemical potential  $\mu$ , we find

$$n = g \left( \frac{mT}{2\pi} \right)^{3/2} e^{-(m-\mu)/T}, \quad (59)$$

$$\rho = m n, \quad (60)$$

$$p = n T \ll \rho. \quad (61)$$

The average energy density per particle is

$$\langle E \rangle \equiv \frac{\rho}{n} = m + \frac{3}{2} T. \quad (62)$$

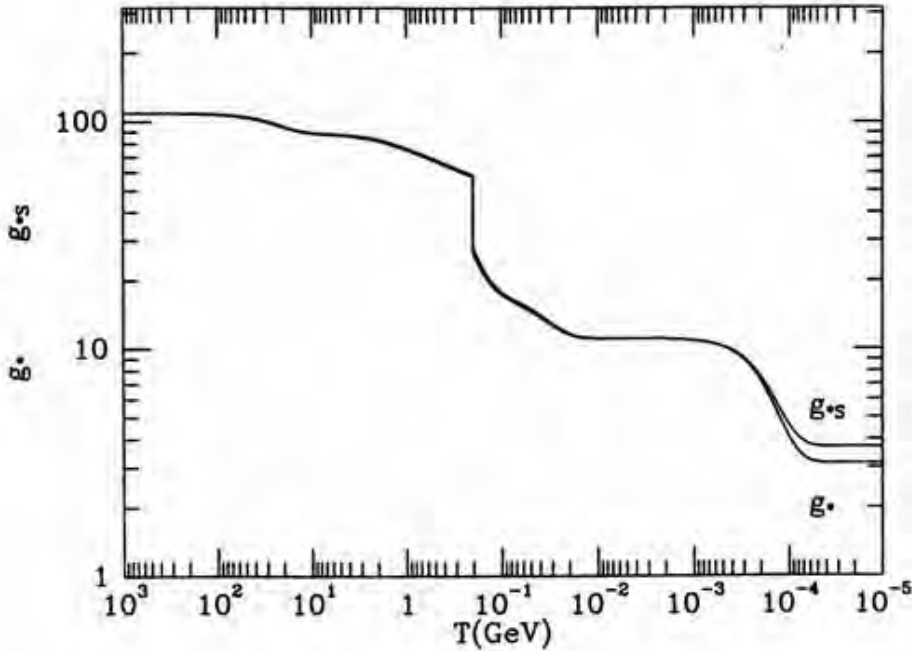
Note that, at any given temperature  $T$ , the contribution to the energy density of the universe coming from non-relativistic particles in thermal equilibrium is exponentially suppressed with respect to that of relativistic particles, therefore we can write

$$\rho_R = \frac{\pi^2}{30} g_* T^4, \quad p_R = \frac{1}{3} \rho_R, \quad (63)$$

$$g_*(T) = \sum_{\text{bosons}} g_i \left( \frac{T_i}{T} \right)^4 + \frac{7}{8} \sum_{\text{fermions}} g_i \left( \frac{T_i}{T} \right)^4, \quad (64)$$

where the factor  $7/8$  takes into account the difference between the Fermi and Bose statistics;  $g_*$  is the total number of light d.o.f. ( $m \ll T$ ), and we have also considered the possibility that particle species  $i$  (bosons or fermions) have an equilibrium distribution at a temperature  $T_i$  different from that of photons, as happens for example when a given relativistic species decouples from the thermal bath, as we will discuss later. This number,  $g_*$ , strongly depends on the temperature of the universe, since as it expands

and cools, different particles go out of equilibrium or become non-relativistic ( $m \gg T$ ) and thus become exponentially suppressed from that moment on. A plot of the time evolution of  $g_*(T)$  can be seen in Fig. 6. For example, for  $T \ll 1$  MeV, i.e. after the time of primordial Big Bang Nucleosynthesis (BBN) and neutrino decoupling, the only relativistic species are the 3 light neutrinos and the photons; since the temperature of the neutrinos is  $T_\nu = (4/11)^{1/3}T_\gamma = 1.90$  K, see below, we have  $g_* = 2 + 3 \times \frac{7}{4} \times \left(\frac{4}{11}\right)^{4/3} = 3.36$ , while  $g_{*S} = 2 + 3 \times \frac{7}{4} \times \left(\frac{4}{11}\right) = 3.91$ .



**Fig. 6:** The light degrees of freedom  $g_*$  and  $g_{*S}$  as a function of the temperature of the universe. From Ref. [16].

For  $1 \text{ MeV} \ll T \ll 100 \text{ MeV}$ , i.e. between BBN and the phase transition from a quark-gluon plasma to hadrons and mesons, we have, as relativistic species, apart from neutrinos and photons, also the electrons and positrons, so  $g_* = 2 + 3 \times \frac{7}{4} + 2 \times \frac{7}{4} = 10.75$ .

For  $T \gg 250 \text{ GeV}$ , i.e. above the electroweak (EW) symmetry breaking scale, we have one photon (2 polarizations), 8 gluons (massless), the  $W^\pm$  and  $Z^0$  (massive), 3 families of quarks & leptones, a Higgs (still undiscovered), with which one finds  $g_* = \frac{427}{4} = 106.75$ .

At temperatures well above the electroweak transition we ignore the number of d.o.f. of particles, since we have never explored those energies in particle physics accelerators. Perhaps in the near future, with the results of the Large Hadron Collider (LHC) at CERN, we may predict the behaviour of the universe at those energy scales. For the moment we even ignore whether the universe was in thermal equilibrium at those temperatures. The highest energy scale at which we can safely say the universe was in thermal equilibrium is that of BBN, i.e. 1 MeV, due to the fact that we observe the present relative abundances of the light element produced at that time. For instance, we can't even claim that the universe went through the quark-gluon phase transition, at  $\sim 200 \text{ MeV}$ , since we have not observed yet any signature of such an event, not to mention the electroweak phase transition, at  $\sim 1 \text{ TeV}$ .

Let us now use the relation between the rate of expansion and the temperature of relativistic particles to obtain the time scale of the universe as a function of its temperature,

$$H = 1.66 g_*^{1/2} \frac{T^2}{M_P} = \frac{1}{2t} \quad \implies \quad t = 0.301 g_*^{-1/2} \frac{M_P}{T^2} \sim \left(\frac{T}{\text{MeV}}\right)^{-2} \text{ s}, \quad (65)$$

thus, e.g. at the EW scale (100 GeV) the universe was just  $10^{-10}$  s old, while during the primordial BBN (1 – 0.1 MeV), it was 1 s to 3 min old.

### 2.5.2 The entropy of the universe

During most of the history of the universe, the rates of reaction,  $\Gamma_{\text{int}}$ , of particles in the thermal bath are much bigger than the rate of expansion of the universe,  $H$ , so that local thermal equilibrium was maintained. In this case, the entropy per comoving volume remained constant. In an expanding universe, the second law of thermodynamics, applied to the element of comoving volume, of unit coordinate volume and physical volume  $V = a^3$ , can be written as, see (42),

$$T dS = d(\rho V) + p dV = d[(\rho + p)V] - V dp. \quad (66)$$

Using the Maxwell condition of integrability,  $\frac{\partial^2 S}{\partial T \partial V} = \frac{\partial^2 S}{\partial V \partial T}$ , we find that  $dp = (\rho + p)dT/T$ , so that

$$dS = d \left[ (\rho + p) \frac{V}{T} + \text{const.} \right], \quad (67)$$

i.e. the entropy in a comoving volume is  $S = (\rho + p)V/T$ , except for a constant. Using now the first law, the covariant conservation of energy  $T^{\mu\nu}_{;\nu} = 0$ , we have

$$d[(\rho + p)a^3] = a^3 dp \quad \implies \quad d\left((\rho + p)\frac{a^3}{T}\right) = 0, \quad (68)$$

and thus, in thermal equilibrium, the total entropy in a comoving volume,  $S = a^3(\rho + p)/T$ , is conserved. During most of the evolution of the universe, this entropy was dominated by the contribution from relativistic particles,

$$S = \frac{2\pi^2}{45} g_{*S} (aT)^3 = \text{const.}, \quad (69)$$

$$g_{*S}(T) = \sum_{\text{bosons}} g_i \left(\frac{T_i}{T}\right)^3 + \frac{7}{8} \sum_{\text{fermions}} g_i \left(\frac{T_i}{T}\right)^3, \quad (70)$$

where  $g_{*S}$  is the number of “entropic” degrees of freedom, as we can see in Fig. 8. Above the electron-positron annihilation, all relativistic particles had the same temperature and thus  $g_{*S} = g_*$ . It may be also useful to realize that the entropy density,  $s = S/a^3$ , is proportional to the number density of relativistic particles, and in particular to the number density of photons,  $s = 1.80g_{*S} n_\gamma$ ; today,  $s = 7.04 n_\gamma$ . However, since  $g_{*S}$  in general is a function of temperature, we can’t always interchange  $s$  and  $n_\gamma$ .

The conservation of  $S$  implies that the entropy density satisfies  $s \propto a^{-3}$ , and thus the physical size of the comoving volume is  $a^3 \propto s^{-1}$ ; therefore, the number of particles of a given species in a comoving volume,  $N = a^3 n$ , is proportional to the number density of that species over the entropy density  $s$ ,

$$N \sim \frac{n}{s} = \begin{cases} \frac{45\zeta(3)g}{2\pi^4 g_{*S}} & T \gg m, \mu \\ \frac{45g}{4\pi^5 \sqrt{2} g_{*S}} \left(\frac{m}{T}\right)^{3/2} e^{-\frac{m-\mu}{T}} & T \ll m \end{cases} \quad (71)$$

If this number does not change, i.e. if those particles are neither created nor destroyed, then  $n/s$  remains constant. As a useful example, we will consider the barionic number in a comoving volume,

$$\frac{n_B}{s} \equiv \frac{n_b - n_{\bar{b}}}{s}. \quad (72)$$

As long as the interactions that violate baryon number occur sufficiently slowly, the baryonic number per comoving volume,  $n_B/s$ , will remain constant. Although

$$\eta \equiv \frac{n_B}{n_\gamma} = 1.80 g_{*S} \frac{n_B}{s}, \quad (73)$$

the ratio between baryon and photon numbers it does not remain constant during the whole evolution of the universe since  $g_{*S}$  varies; e.g. during the annihilation of electrons and positrons, the number of photons per comoving volume,  $N_\gamma = a^3 n_\gamma$ , grows a factor 11/4, and  $\eta$  decreases by the same factor. After this epoch, however,  $g_*$  is constant so that  $\eta \simeq 7n_B/s$  and  $n_B/s$  can be used indistinctly.

Another consequence of Eq. (69) is that  $S = \text{const.}$  implies that the temperature of the universe evolves as

$$T \propto g_{*S}^{-1/3} a^{-1}. \quad (74)$$

As long as  $g_{*S}$  remains constant, we recover the well known result that the universe cools as it expands according to  $T \propto 1/a$ . The factor  $g_{*S}^{-1/3}$  appears because when a species becomes non-relativistic (when  $T \leq m$ ), and effectively disappears from the energy density of the universe, its entropy is transferred to the rest of the relativistic particles in the plasma, making  $T$  decrease not as quickly as  $1/a$ , until  $g_{*S}$  again becomes constant.

From the observational fact that the universe expands today one can deduce that in the past it must have been hotter and denser, and that in the future it will be colder and more dilute. Since the ratio of scale factors is determined by the redshift parameter  $z$ , we can obtain (to very good approximation) the temperature of the universe in the past with

$$T = T_0 (1 + z). \quad (75)$$

This expression has been spectacularly confirmed thanks to the absorption spectra of distant quasars [17]. These spectra suggest that the radiation background was acting as a thermal bath for the molecules in the interstellar medium with a temperature of 9 K at a redshift  $z \sim 2$ , and thus that in the past the photon background was hotter than today. Furthermore, observations of the anisotropies in the microwave background confirm that the universe at a redshift  $z = 1089$  had a temperature of 0.3 eV, in agreement with Eq. (75).

## 2.6 The thermal evolution of the universe

In a strict mathematical sense, it is impossible for the universe to have been always in thermal equilibrium since the FRW model does not have a timelike Killing vector. In practice, however, we can say that the universe has been most of its history very close to thermal equilibrium. Of course, those periods in which there were deviations from thermal equilibrium have been crucial for its evolution thereafter (e.g. baryogenesis, QCD transition, primordial nucleosynthesis, recombination, etc.); without these the universe today would be very different and probably we would not be here to tell the story.

The key to understand the thermal history of the universe is the comparison between the rates of interaction between particles (microphysics) and the rate of expansion of the universe (macrophysics). Ignoring for the moment the dependence of  $g_*$  on temperature, the rate of change of  $T$  is given directly by the rate of expansion,  $\dot{T}/T = -H$ . As long as the local interactions – necessary in order that the particle distribution function adjusts *adiabatically* to the change of temperature – are sufficiently fast compared with the rate of expansion of the universe, the latter will evolve as a succession of states very close to thermal equilibrium, with a temperature proportional to  $a^{-1}$ . If we evaluate the interaction rates as

$$\Gamma_{\text{int}} \equiv \langle n \sigma |v| \rangle, \quad (76)$$

where  $n(t)$  is the number density of target particles,  $\sigma$  is the cross section on the interaction and  $v$  is the relative velocity of the reaction, all averaged on a thermal distribution; then a rule of thumb for ensuring



that thermal equilibrium is maintained is

$$\Gamma_{\text{int}} \gtrsim H. \quad (77)$$

This criterium is understandable. Suppose, as often occurs, that the interaction rate in thermal equilibrium is  $\Gamma_{\text{int}} \propto T^n$ , with  $n > 2$ ; then, the number of interactions of a particle after time  $t$  is

$$N_{\text{int}} = \int_t^\infty \Gamma_{\text{int}}(t') dt' = \frac{1}{n-2} \frac{\Gamma_{\text{int}}}{H}(t), \quad (78)$$

therefore the particle interacts less than once from the moment in which  $\Gamma_{\text{int}} \approx H$ . If  $\Gamma_{\text{int}} \gtrsim H$ , the species remains coupled to the thermal plasma. This doesn't mean that, necessarily, the particle is out of local thermal equilibrium, since we have seen already that relativistic particles that have decoupled retain their equilibrium distribution, only at a different temperature from that of the rest of the plasma.

In order to obtain an approximate description of the decoupling of a particle species in an expanding universe, let us consider two types of interaction:

i) interactions mediated by massless gauge bosons, like for example the photon. In this case, the cross section for particles with significant momentum transfer can be written as  $\sigma \sim \alpha^2/T^2$ , with  $\alpha = g^2/4\pi$  the coupling constant of the interaction. Assuming local thermal equilibrium,  $n(t) \sim T^3$  and thus the interaction rate becomes  $\Gamma \sim n \sigma |v| \sim \alpha^2 T$ . Therefore,

$$\frac{\Gamma}{H} \sim \alpha^2 \frac{M_P}{T}, \quad (79)$$

so that for temperatures of the universe  $T \lesssim \alpha^2 M_P \sim 10^{16}$  GeV, the reactions are fast enough and the plasma is in equilibrium, while for  $T \gtrsim 10^{16}$  GeV, reactions are too slow to maintain equilibrium and it is said that they are ‘‘frozen-out’’. An important consequence of this result is that the universe could never have been in thermal equilibrium above the grand unification (GUT) scale.

ii) interactions mediated by massive gauge bosons, e.g. like the  $W^\pm$  and  $Z^0$ , or those responsible for the GUT interactions,  $X$  and  $Y$ . We will generically call them  $X$  bosons. The cross section depends rather strongly on the temperature of the plasma,

$$\sigma \sim \begin{cases} G_X^2 T^2 & T \ll M_X \\ \frac{\alpha^2}{T^2} & T \gg M_X \end{cases} \quad (80)$$

where  $G_X \sim \alpha/M_X^2$  is the effective coupling constant of the interaction at energies well below the mass of the vector boson, analogous to the Fermi constant of the electroweak interaction,  $G_F = g^2/(4\sqrt{2}M_W^2)$  at tree level. Note that for  $T \gg M_X$  we recover the result for massless bosons, so we will concentrate here on the other case. For  $T \leq M_X$ , the rate of thermal interactions is  $\Gamma \sim n \sigma |v| \sim G_X^2 T^5$ . Therefore,

$$\frac{\Gamma}{H} \sim G_X^2 M_P T^3, \quad (81)$$

such that at temperatures in the range

$$M_X \gtrsim T \gtrsim G_X^{-2/3} M_P^{-1/3} \sim \left( \frac{M_X}{100 \text{ GeV}} \right)^{4/3} \text{ MeV}, \quad (82)$$

reactions occur so fast that the plasma is in thermal equilibrium, while for  $T \lesssim (M_X/100 \text{ GeV})^{4/3} \text{ MeV}$ , those reactions are too slow for maintaining equilibrium and they effective freeze-out, see Eq. (78).

### 2.6.1 The decoupling of relativistic particles

Those relativistic particles that have decoupled from the thermal bath do not participate in the transfer of entropy when the temperature of the universe falls below the mass threshold of a given species  $T \simeq m$ ; in fact, the temperature of the decoupled relativistic species falls as  $T \propto 1/a$ , as we will now show. Suppose that a relativistic particle is initially in local thermal equilibrium, and that it decouples at a temperature  $T_D$  and time  $t_D$ . The phase space distribution at the time of decoupling is given by the equilibrium distribution,

$$f(\mathbf{p}, t_D) = \frac{1}{e^{E/T_D} \pm 1}. \quad (83)$$

After decoupling, the energy of each massless particle suffers redshift,  $E(t) = E_D (a_D/a(t))$ . The number density of particles also decreases,  $n(t) = n_D (a_D/a(t))^3$ . Thus, the phase space distribution at a time  $t > t_D$  is

$$f(\mathbf{p}, t) = \frac{d^3 n}{d^3 \mathbf{p}} = f\left(\mathbf{p} \frac{a}{a_D}, t_D\right) = \frac{1}{e^{E a/a_D T_D} \pm 1} = \frac{1}{e^{E/T} \pm 1}, \quad (84)$$

so that we conclude that the distribution function of a particle that has decoupled while being relativistic remains self-similar as the universe expands, with a temperature that decreases as

$$T = T_D \frac{a_D}{a} \propto a^{-1}, \quad (85)$$

and *not* as  $g_{*S}^{-1/3} a^{-1}$ , like the rest of the plasma in equilibrium (74).

### 2.6.2 The decoupling of non-relativistic particles

Those particles that decoupled from the thermal bath when they were non-relativistic ( $m \gg T$ ) behave differently. Let us study the evolution of the distribution function of a non-relativistic particle that was in local thermal equilibrium at a time  $t_D$ , when the universe had a temperature  $T_D$ . The moment of each particle suffers redshift as the universe expands,  $|\mathbf{p}| = |\mathbf{p}_D| (a_D/a)$ , see Eq. (24). Therefore, their kinetic energy satisfies  $E = E_D (a_D/a)^2$ . On the other hand, the particle number density also varies,  $n(t) = n_D (a_D/a(t))^3$ , so that a decoupled non-relativistic particle will have an equilibrium distribution function characterized by a temperature

$$T = T_D \frac{a_D^2}{a^2} \propto a^{-2}, \quad (86)$$

and a chemical potential

$$\mu(t) = m + (\mu_D - m) \frac{T}{T_D}, \quad (87)$$

whose variation is precisely that which is needed for the number density of particle to decrease as  $a^{-3}$ .

In summary, a particle species that decouples from the thermal bath follows an equilibrium distribution function with a temperature that decreases like  $T_R \propto a^{-1}$  for relativistic particles ( $T_D \gg m$ ) or like  $T_{NR} \propto a^{-2}$  for non-relativistic particles ( $T_D \ll m$ ). On the other hand, for semi-relativistic particles ( $T_D \sim m$ ), its phase space distribution *does not maintain* an equilibrium distribution function, and should be computed case by case.

### 2.6.3 Brief thermal history of the universe

I will briefly summarize here the thermal history of the universe, from the Planck era to the present. As we go back in time, the universe becomes hotter and hotter and thus the amount of energy available for particle interactions increases. As a consequence, the nature of interactions goes from those described at low energy by long range gravitational and electromagnetic physics, to atomic physics, nuclear physics,

all the way to high energy physics at the electroweak scale, grand unification (perhaps), and finally quantum gravity. The last two are still uncertain since we do not have any experimental evidence for those ultra high energy phenomena, and perhaps Nature has followed a different path.

The way we know about the high energy interactions of matter is via particle accelerators, which are unravelling the details of those fundamental interactions as we increase in energy. However, one should bear in mind that the physical conditions that take place in our high energy colliders are very different from those that occurred in the early universe. These machines could never reproduce the conditions of density and pressure in the rapidly expanding thermal plasma of the early universe. Nevertheless, those experiments are crucial in understanding the nature and *rate* of the local fundamental interactions available at those energies. What interests cosmologists is the statistical and thermal properties that such a plasma should have, and the role that causal horizons play in the final outcome of the early universe expansion. For instance, of crucial importance is the time at which certain particles *decoupled* from the plasma, i.e. when their interactions were not quick enough compared with the expansion of the universe, and they were left out of equilibrium with the plasma.

One can trace the evolution of the universe from its origin till today. There is still some speculation about the physics that took place in the universe above the energy scales probed by present colliders. Nevertheless, the overall layout presented here is a plausible and hopefully testable proposal. According to the best accepted view, the universe must have originated at the Planck era ( $10^{19}$  GeV,  $10^{-43}$  s) from a quantum gravity fluctuation. Needless to say, we don't have any experimental evidence for such a statement: Quantum gravity phenomena are still in the realm of physical speculation. However, it is plausible that a primordial era of cosmological *inflation* originated then. Its consequences will be discussed below. Soon after, the universe may have reached the Grand Unified Theories (GUT) era ( $10^{16}$  GeV,  $10^{-35}$  s). Quantum fluctuations of the inflaton field most probably left their imprint then as tiny perturbations in an otherwise very homogenous patch of the universe. At the end of inflation, the huge energy density of the inflaton field was converted into particles, which soon thermalized and became the origin of the hot Big Bang as we know it. Such a process is called *reheating* of the universe. Since then, the universe became radiation dominated. It is probable (although by no means certain) that the asymmetry between matter and antimatter originated at the same time as the rest of the energy of the universe, from the decay of the inflaton. This process is known under the name of *baryogenesis* since baryons (mostly quarks at that time) must have originated then, from the leftovers of their annihilation with antibaryons. It is a matter of speculation whether baryogenesis could have occurred at energies as low as the electroweak scale (100 GeV,  $10^{-10}$  s). Note that although particle physics experiments have reached energies as high as 100 GeV, we still do not have observational evidence that the universe actually went through the EW phase transition. If confirmed, baryogenesis would constitute another "window" into the early universe. As the universe cooled down, it may have gone through the quark-gluon phase transition ( $10^2$  MeV,  $10^{-5}$  s), when baryons (mainly protons and neutrons) formed from their constituent quarks.

The furthest window we have on the early universe at the moment is that of *primordial nucleosynthesis* (1 – 0.1 MeV, 1 s – 3 min), when protons and neutrons were cold enough that bound systems could form, giving rise to the lightest elements, soon after *neutrino decoupling*: It is the realm of nuclear physics. The observed relative abundances of light elements are in agreement with the predictions of the hot Big Bang theory. Immediately afterwards, electron-positron annihilation occurs (0.5 MeV, 1 min) and all their energy goes into photons. Much later, at about (1 eV,  $\sim 10^5$  yr), matter and radiation have equal energy densities. Soon after, electrons become bound to nuclei to form atoms (0.3 eV,  $3 \times 10^5$  yr), in a process known as *recombination*: It is the realm of atomic physics. Immediately after, photons decouple from the plasma, travelling freely since then. Those are the photons we observe as the cosmic microwave background. Much later ( $\sim 1 - 10$  Gyr), the small inhomogeneities generated during inflation have grown, via gravitational collapse, to become galaxies, clusters of galaxies, and superclusters, characterizing the epoch of *structure formation*. It is the realm of long range gravitational physics, perhaps

dominated by a vacuum energy in the form of a cosmological constant. Finally (3K, 13 Gyr), the Sun, the Earth, and biological life originated from previous generations of stars, and from a primordial soup of organic compounds, respectively.

I will now review some of the more robust features of the Hot Big Bang theory of which we have precise observational evidence.

#### 2.6.4 Primordial nucleosynthesis and light element abundance

In this subsection I will briefly review Big Bang nucleosynthesis and give the present observational constraints on the amount of baryons in the universe. In 1920 Eddington suggested that the sun might derive its energy from the fusion of hydrogen into helium. The detailed reactions by which stars burn hydrogen were first laid out by Hans Bethe in 1939. Soon afterwards, in 1946, George Gamow realized that similar processes might have occurred also in the hot and dense early universe and gave rise to the first light elements [4]. These processes could take place when the universe had a temperature of around  $T_{\text{NS}} \sim 1 - 0.1$  MeV, which is about 100 times the temperature in the core of the Sun, while the density is  $\rho_{\text{NS}} = \frac{\pi^2}{30} g_* T_{\text{NS}}^4 \sim 82 \text{ g cm}^{-3}$ , about the same density as the core of the Sun. Note, however, that although both processes are driven by identical thermonuclear reactions, the physical conditions in star and Big Bang nucleosynthesis are very different. In the former, gravitational collapse heats up the core of the star and reactions last for billions of years (except in supernova explosions, which last a few minutes and creates all the heavier elements beyond iron), while in the latter the universe expansion cools the hot and dense plasma in just a few minutes. Nevertheless, Gamow reasoned that, although the early period of cosmic expansion was much shorter than the lifetime of a star, there was a large number of free neutrons at that time, so that the lighter elements could be built up quickly by successive neutron captures, starting with the reaction  $n + p \rightarrow D + \gamma$ . The abundances of the light elements would then be correlated with their neutron capture cross sections, in rough agreement with observations [6, 18].

Nowadays, Big Bang nucleosynthesis (BBN) codes compute a chain of around 30 coupled nuclear reactions [19], to produce all the light elements up to beryllium-7.<sup>3</sup> Only the first four or five elements can be computed with accuracy better than 1% and compared with cosmological observations. These light elements are  $H$ ,  ${}^4\text{He}$ ,  $D$ ,  ${}^3\text{He}$ ,  ${}^7\text{Li}$ , and perhaps also  ${}^6\text{Li}$ . Their observed relative abundance to hydrogen is  $[1 : 0.25 : 3 \cdot 10^{-5} : 2 \cdot 10^{-5} : 2 \cdot 10^{-10}]$  with various errors, mainly systematic. The BBN codes calculate these abundances using the laboratory measured nuclear reaction rates, the decay rate of the neutron, the number of light neutrinos and the homogeneous FRW expansion of the universe, as a function of *only* one variable, the number density fraction of baryons to photons,  $\eta \equiv n_{\text{B}}/n_{\gamma}$ . In fact, the present observations are only consistent, see Fig. 9 and Ref. [18–20], with a very narrow range of values of

$$\eta_{10} \equiv 10^{10} \eta = 6.2 \pm 0.6. \quad (88)$$

Such a small value of  $\eta$  indicates that there is about one baryon per  $10^9$  photons in the universe today. Any acceptable theory of baryogenesis should account for such a small number. Furthermore, the present baryon fraction of the critical density can be calculated from  $\eta_{10}$  as

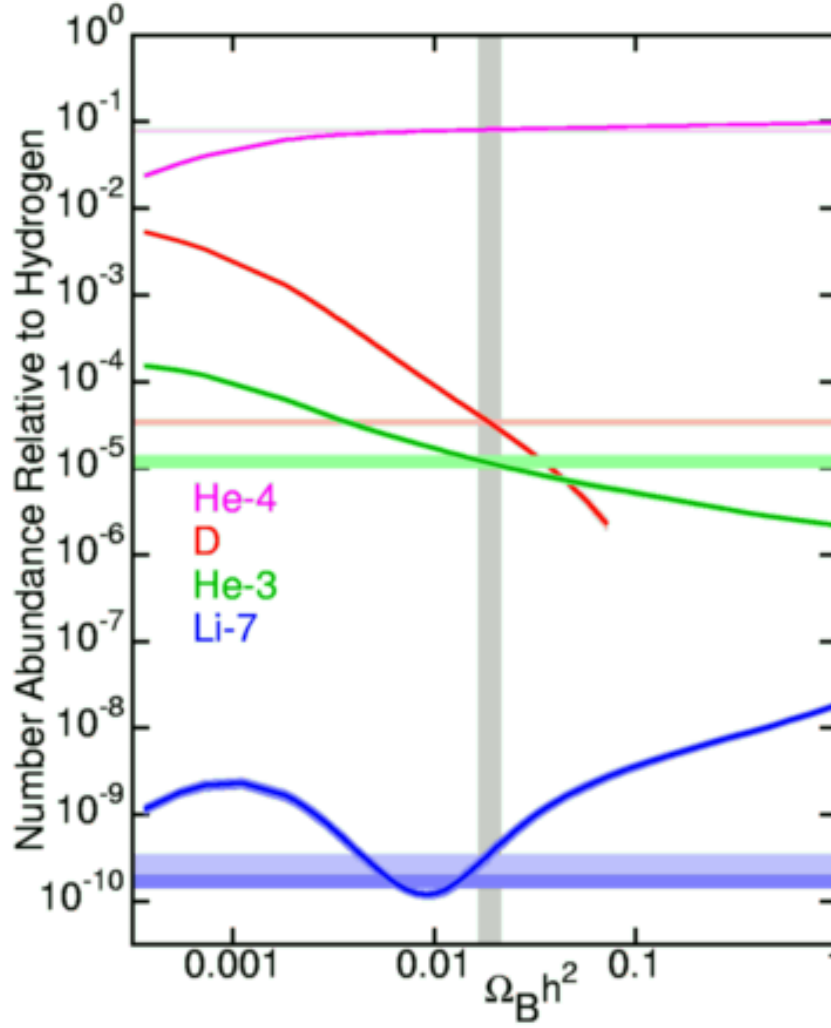
$$\Omega_{\text{B}} h^2 = 3.6271 \times 10^{-3} \eta_{10} = 0.0224 \pm 0.0022 \quad (95\% \text{ c.l.}) \quad (89)$$

Clearly, this number is well below closure density, so baryons cannot account for all the matter in the universe, as I shall discuss below.

#### 2.6.5 Neutrino decoupling

Just before the nucleosynthesis of the lightest elements in the early universe, weak interactions were too slow to keep neutrinos in thermal equilibrium with the plasma, so they decoupled. We can estimate the

<sup>3</sup>The rest of nuclei, up to iron (Fe), are produced in heavy stars, and beyond Fe in novae and supernovae explosions.



**Fig. 7:** The relative abundance of light elements to Hydrogen. Note the large range of scales involved. From Ref. [18].

temperature at which decoupling occurred from the weak interaction cross section,  $\sigma_w \simeq G_F^2 T^2$  at finite temperature  $T$ , where  $G_F = 1.2 \times 10^{-5} \text{ GeV}^{-2}$  is the Fermi constant. The neutrino interaction rate, via W boson exchange in  $n + \nu \leftrightarrow p + e^-$  and  $p + \bar{\nu} \leftrightarrow n + e^+$ , can be written as [16]

$$\Gamma_\nu = n_\nu \langle \sigma_w |v| \rangle \simeq G_F^2 T^5, \quad (90)$$

while the rate of expansion of the universe at that time ( $g_* = 10.75$ ) was  $H \simeq 5.4 T^2/M_P$ , where  $M_P = 1.22 \times 10^{19} \text{ GeV}$  is the Planck mass. Neutrinos decouple when their interaction rate is slower than the universe expansion,  $\Gamma_\nu \leq H$  or, equivalently, at  $T_{\nu\text{-dec}} \simeq 0.8 \text{ MeV}$ . Below this temperature, neutrinos are no longer in thermal equilibrium with the rest of the plasma, and their temperature continues to decay inversely proportional to the scale factor of the universe. Since neutrinos decoupled before  $e^+e^-$  annihilation, the cosmic background of neutrinos has a temperature today lower than that of the microwave background of photons. Let us compute the difference. At temperatures above the mass of the electron,  $T > m_e = 0.511 \text{ MeV}$ , and below  $0.8 \text{ MeV}$ , the only particle species contributing to the entropy of the universe are the photons ( $g_* = 2$ ) and the electron-positron pairs ( $g_* = 4 \times \frac{7}{8}$ ); total number of degrees of freedom  $g_* = \frac{11}{2}$ . At temperatures  $T \simeq m_e$ , electrons and positrons annihilate into photons, heating up the plasma (but not the neutrinos, which had decoupled already). At temperatures

$T < m_e$ , only photons contribute to the entropy of the universe, with  $g_* = 2$  degrees of freedom. Therefore, from the conservation of entropy, we find that the ratio of  $T_\gamma$  and  $T_\nu$  today must be

$$\frac{T_\gamma}{T_\nu} = \left(\frac{11}{4}\right)^{1/3} = 1.401 \quad \Rightarrow \quad T_\nu = 1.945 \text{ K}, \quad (91)$$

where I have used  $T_{\text{CMB}} = 2.725 \pm 0.002 \text{ K}$ . We still have not measured such a relic background of neutrinos, and probably will remain undetected for a long time, since they have an average energy of order  $10^{-4} \text{ eV}$ , much below that required for detection by present experiments (of order GeV), precisely because of the relative weakness of the weak interactions. Nevertheless, it would be fascinating if, in the future, ingenious experiments were devised to detect such a background, since it would confirm one of the most robust features of Big Bang cosmology.

### 2.6.6 Matter-radiation equality

Relativistic species have energy densities proportional to the quartic power of temperature and therefore scale as  $\rho_R \propto a^{-4}$ , while non-relativistic particles have essentially zero pressure and scale as  $\rho_M \propto a^{-3}$ . Therefore, there will be a time in the evolution of the universe in which both energy densities are equal  $\rho_R(t_{\text{eq}}) = \rho_M(t_{\text{eq}})$ . Since then both decay differently, and thus

$$1 + z_{\text{eq}} = \frac{a_0}{a_{\text{eq}}} = \frac{\Omega_M}{\Omega_R} = 3.1 \times 10^4 \Omega_M h^2, \quad (92)$$

where I have used  $\Omega_R h^2 = \Omega_{\text{CMB}} h^2 + \Omega_\nu h^2 = 3.24 \times 10^{-5}$  for three massless neutrinos at  $T = T_\nu$ . As I will show later, the matter content of the universe today is below critical,  $\Omega_M \simeq 0.3$ , while  $h \simeq 0.71$ , and therefore  $(1 + z_{\text{eq}}) \simeq 3400$ , or about  $t_{\text{eq}} = 1308 (\Omega_M h^2)^{-2} \text{yr} \simeq 61,000$  years after the origin of the universe. Around the time of matter-radiation equality, the rate of expansion (19) can be written as ( $a_0 \equiv 1$ )

$$H(a) = H_0 \left( \Omega_R a^{-4} + \Omega_M a^{-3} \right)^{1/2} = H_0 \Omega_M^{1/2} a^{-3/2} \left( 1 + \frac{a_{\text{eq}}}{a} \right)^{1/2}. \quad (93)$$

The *horizon size* is the coordinate distance travelled by a photon since the beginning of the universe,  $d_H \sim H^{-1}$ , i.e. the size of causally connected regions in the universe. The *comoving* horizon size is then given by

$$d_H(a) = \int \frac{da}{a^2 H(a)} = \frac{2c H_0^{-1}}{\sqrt{\Omega_M(1 + z_{\text{eq}})}} \left( \sqrt{\frac{a}{a_{\text{eq}}}} + 1 - 1 \right). \quad (94)$$

Thus the horizon size at matter-radiation equality ( $a = a_{\text{eq}}$ ) is

$$d_H(a_{\text{eq}}) \simeq 14 (\Omega_M h)^{-1} \text{Mpc}/h. \quad (95)$$

As we will see later, this scale plays a very important role in theories of structure formation.

### 2.6.7 Recombination and photon decoupling

As the temperature of the universe decreased, electrons could eventually become bound to protons to form neutral hydrogen. Nevertheless, there is always a non-zero probability that a rare energetic photon ionizes hydrogen and produces a free electron. The *ionization fraction* of electrons in equilibrium with the plasma at a given temperature is given by the Saha equation [16]

$$\frac{1 - X_e^{\text{eq}}}{(X_e^{\text{eq}})^2} = \frac{4\sqrt{2}\zeta(3)}{\sqrt{\pi}} \eta \left( \frac{T}{m_e} \right)^{3/2} e^{E_{\text{ion}}/T}, \quad (96)$$

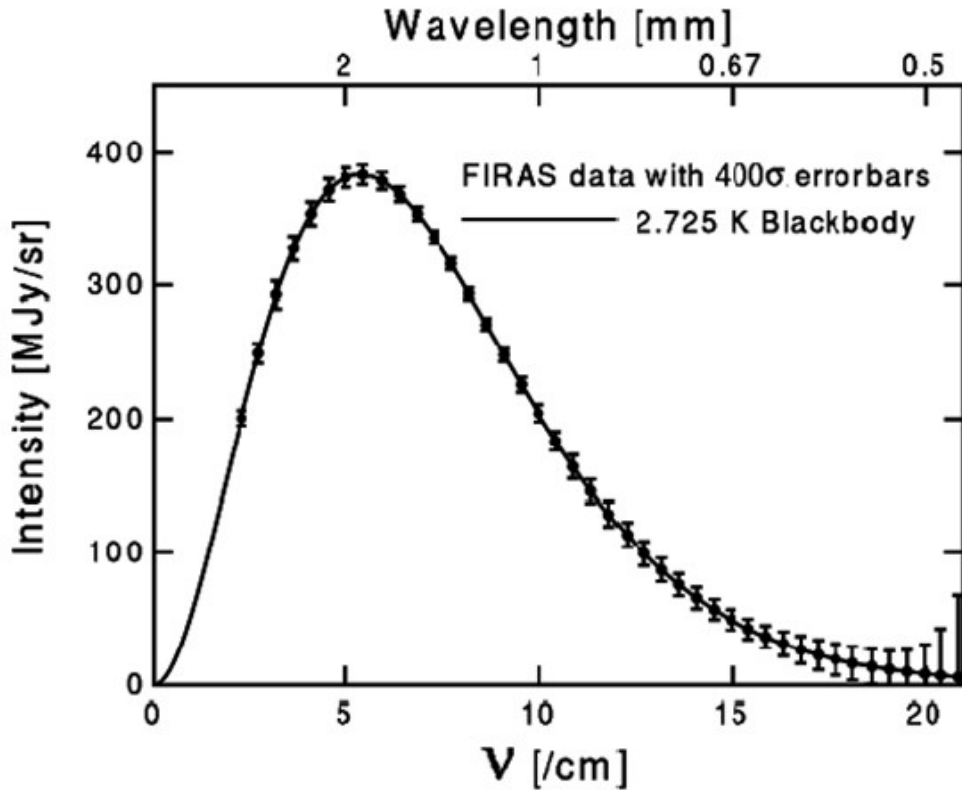
where  $E_{\text{ion}} = 13.6 \text{ eV}$  is the ionization energy of hydrogen, and  $\eta$  is the baryon-to-photon ratio (88). If we now use Eq. (75), we can compute the ionization fraction  $X_e^{\text{eq}}$  as a function of redshift  $z$ . Note that

the huge number of photons with respect to electrons (in the ratio  ${}^4\text{He} : \text{H} : \gamma \simeq 1 : 4 : 10^{10}$ ) implies that even at a very low temperature, the photon distribution will contain a sufficiently large number of high-energy photons to ionize a significant fraction of hydrogen. In fact, *defining* recombination as the time at which  $X_e^{\text{eq}} \equiv 0.1$ , one finds that the recombination temperature is  $T_{\text{rec}} = 0.296 \text{ eV} \ll E_{\text{ion}}$ , for  $\eta_{10} \simeq 6.1$ . Comparing with the present temperature of the microwave background, we deduce the corresponding redshift at recombination,  $(1 + z_{\text{rec}}) \simeq 1260$ .

Photons remain in thermal equilibrium with the plasma of baryons and electrons through elastic Thomson scattering, with cross section

$$\sigma_T = \frac{8\pi\alpha^2}{3m_e^2} = 6.65 \times 10^{-25} \text{ cm}^2 = 0.665 \text{ barn}, \quad (97)$$

where  $\alpha = 1/137.036$  is the dimensionless electromagnetic coupling constant. The mean free path of photons  $\lambda_\gamma$  in such a plasma can be estimated from the photon interaction rate,  $c\lambda_\gamma^{-1} \simeq \Gamma_\gamma = n_e\sigma_T c$ . For temperatures above a few eV, the mean free path is much smaller than the causal horizon at that time and photons suffer multiple scattering: the plasma is like a dense fog. Photons will decouple from the plasma when their interaction rate cannot keep up with the expansion of the universe and the mean free path becomes larger than the horizon size: the universe becomes transparent. We can estimate this moment by evaluating  $\Gamma_\gamma = H$  at photon decoupling. Using  $n_e = X_e \eta n_\gamma$ , one can compute the decoupling temperature as  $T_{\text{dec}} = 0.256 \text{ eV}$ , and the corresponding redshift as  $1 + z_{\text{dec}} \simeq 1090$ . Recently, WMAP and Planck measured this redshift to be  $1 + z_{\text{dec}} \simeq 1089 \pm 1$  [21]. This redshift defines the so called *last scattering surface*, when photons last scattered off protons and electrons and travelled freely ever since. This decoupling occurred when the universe was approximately  $t_{\text{dec}} = 1.5 \times 10^5 (\Omega_M h^2)^{-1/2} \simeq 380,000$  years old.



**Fig. 8:** The Cosmic Microwave Background Spectrum seen by the FIRAS instrument on COBE. The CMB blackbody spectrum has a temperature  $T_0 = 2.725 \pm 0.002 \text{ K}$ . From Ref. [22].

### 2.6.8 The microwave background

One of the most remarkable observations ever made by mankind is the detection of the relic background of photons from the Big Bang. This background was predicted by George Gamow and collaborators in the 1940s, based on the consistency of primordial nucleosynthesis with the observed helium abundance. They estimated a value of about 10 K, although a somewhat more detailed analysis by Alpher and Herman in 1949 predicted  $T_\gamma \approx 5$  K. Unfortunately, they had doubts whether the radiation would have survived until the present, and this remarkable prediction slipped into obscurity, until Dicke, Peebles, Roll and Wilkinson [23] studied the problem again in 1965. Before they could measure the photon background, they learned that Penzias and Wilson had observed a weak isotropic background signal at a radio wavelength of 7.35 cm, corresponding to a blackbody temperature of  $T_\gamma = 3.5 \pm 1$  K. They published their two papers back to back, with that of Dicke et al. explaining the fundamental significance of their measurement [6].

Since then many different experiments have confirmed the existence of the microwave background. The most outstanding one has been the Cosmic Background Explorer (COBE) satellite, whose FIRAS instrument measured the photon background with great accuracy over a wide range of frequencies ( $\nu = 1 - 97 \text{ cm}^{-1}$ ), see Ref. [22], with a spectral resolution  $\frac{\Delta\nu}{\nu} = 0.0035$ . Nowadays, the photon spectrum is confirmed to be a blackbody spectrum with a temperature given by [22]

$$T_{\text{CMB}} = 2.725 \pm 0.002 \text{ K (systematic, 95\% c.l.)} \pm 7 \mu\text{K (1}\sigma \text{ statistical)} \quad (98)$$

In fact, this is the best blackbody spectrum ever measured, see Fig. 8, with spectral distortions below the level of 10 parts per million (ppm).

Moreover, the differential microwave radiometer (DMR) instrument on COBE, with a resolution of about  $7^\circ$  in the sky, has also confirmed that it is an extraordinarily isotropic background. The deviations from isotropy, i.e. differences in the temperature of the blackbody spectrum measured in different directions in the sky, are of the order of  $20 \mu\text{K}$  on large scales, or one part in  $10^5$ , see Ref. [24]. There is, in fact, a dipole anisotropy of one part in  $10^3$ ,  $\delta T_1 = 3.372 \pm 0.007 \text{ mK}$  (95% c.l.), in the direction of the Virgo cluster,  $(l, b) = (264.14^\circ \pm 0.30, 48.26^\circ \pm 0.30)$  (95% c.l.). Under the assumption that a Doppler effect is responsible for the entire CMB dipole, the velocity of the Sun with respect to the CMB rest frame is  $v_\odot = 371 \pm 0.5 \text{ km/s}$ , see Ref. [22].<sup>4</sup> When subtracted, we are left with a whole spectrum of anisotropies in the higher multipoles (quadrupole, octupole, etc.),  $\delta T_2 = 18 \pm 2 \mu\text{K}$  (95% c.l.), see Ref. [24] and Fig. 9.

Soon after COBE, other groups quickly confirmed the detection of temperature anisotropies at around  $30 \mu\text{K}$  and above, at higher multipole numbers or smaller angular scales. As I shall discuss below, these anisotropies play a crucial role in the understanding of the origin of structure in the universe.

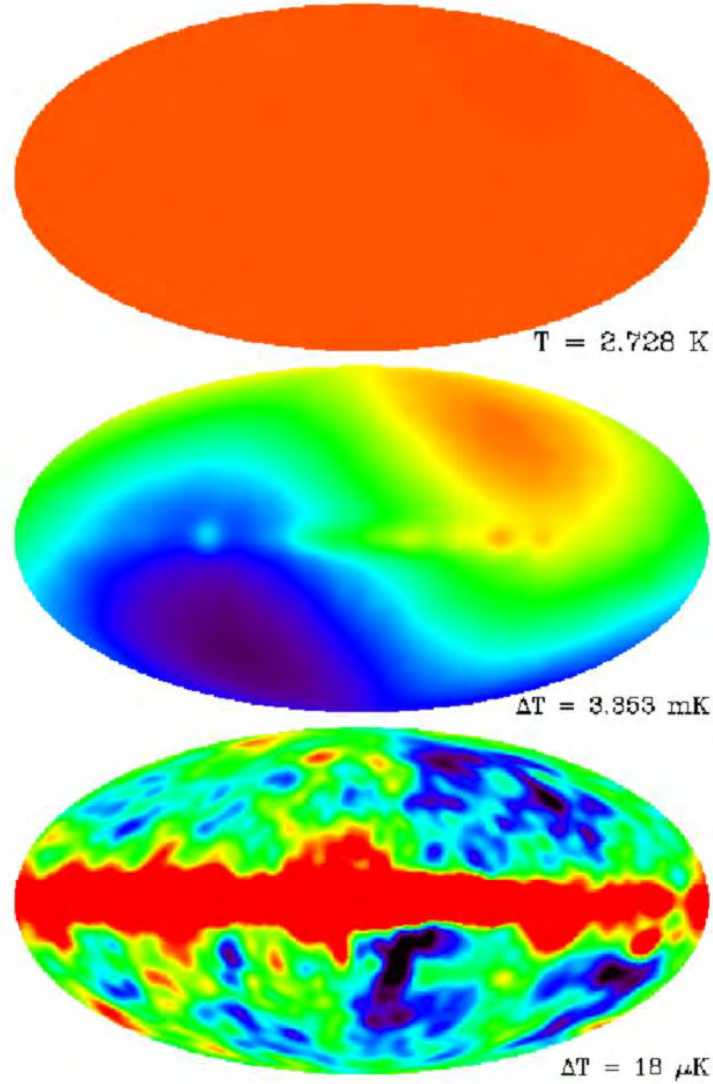
### 2.6.9 Large-scale structure formation

Although the isotropic microwave background indicates that the universe in the *past* was extraordinarily homogeneous, we know that the universe *today* is not exactly homogeneous: we observe galaxies, clusters and superclusters on large scales. These structures are expected to arise from very small primordial inhomogeneities that grow in time via gravitational instability, and that may have originated from tiny ripples in the metric, as matter fell into their troughs. Those ripples must have left some trace as temperature anisotropies in the microwave background, and indeed such anisotropies were finally discovered by the COBE satellite in 1992. The reason why they took so long to be discovered was that they appear as perturbations in temperature of only one part in  $10^5$ .

While the predicted anisotropies have finally been seen in the CMB, not all kinds of matter and/or evolution of the universe can give rise to the structure we observe today. If we define the density contrast

<sup>4</sup>COBE even determined the annual variation due to the Earth's motion around the Sun – the ultimate proof of Copernicus' hypothesis.





**Fig. 9:** The Cosmic Microwave Background Spectrum seen by the DMR instrument on COBE. The top figure corresponds to the monopole,  $T_0 = 2.725 \pm 0.002$  K. The middle figure shows the dipole,  $\delta T_1 = 3.372 \pm 0.014$  mK, and the lower figure shows the quadrupole and higher multipoles,  $\delta T_2 = 18 \pm 2$   $\mu$ K. The central region corresponds to foreground by the galaxy. From Ref. [24].

as [25]

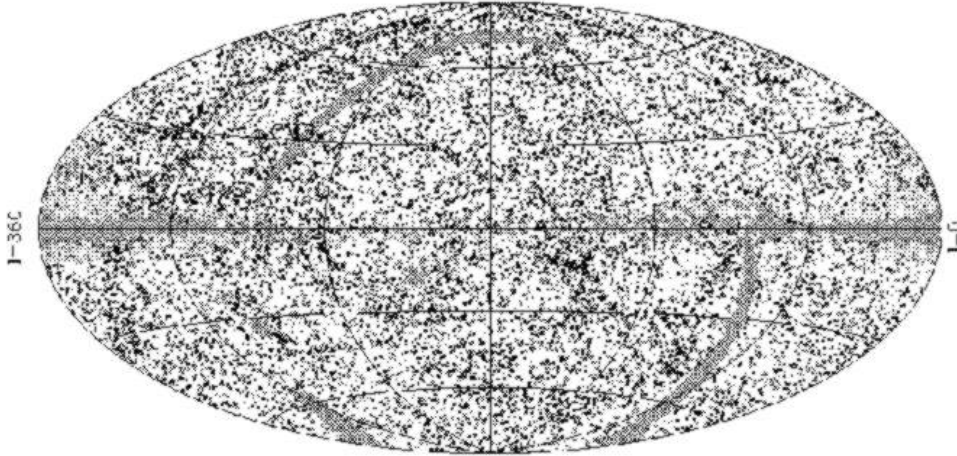
$$\delta(\vec{x}, a) \equiv \frac{\rho(\vec{x}, a) - \bar{\rho}(a)}{\bar{\rho}(a)} = \int d^3\vec{k} \delta_k(a) e^{i\vec{k}\cdot\vec{x}}, \quad (99)$$

where  $\bar{\rho}(a) = \rho_0 a^{-3}$  is the average cosmic density, we need a theory that will grow a density contrast with amplitude  $\delta \sim 10^{-5}$  at the last scattering surface ( $z = 1100$ ) up to density contrasts of the order of  $\delta \sim 10^6$  for galaxies at redshifts  $z \ll 1$ , i.e. today. This is a *necessary* requirement for any consistent theory of structure formation [26].

Furthermore, the anisotropies observed by the Planck satellite correspond to a small-amplitude scale-invariant primordial power spectrum of inhomogeneities

$$P(k) = \langle |\delta_k|^2 \rangle \propto k^n, \quad \text{with } n = 1, \quad (100)$$

where the brackets  $\langle \cdot \rangle$  represent integration over an ensemble of different universe realizations. These inhomogeneities are like waves in the space-time metric. When matter fell in the troughs of those waves, it created density perturbations that collapsed gravitationally to form galaxies and clusters of galaxies, with a spectrum that is also scale invariant. Such a type of spectrum was proposed in the early 1970s by Edward R. Harrison, and independently by the Russian cosmologist Yakov B. Zel'dovich, see Ref. [27], to explain the distribution of galaxies and clusters of galaxies on very large scales in our observable universe.



**Fig. 10:** The IRAS Point Source Catalog redshift survey contains some 15,000 galaxies, covering over 83% of the sky up to redshifts of  $z \leq 0.05$ . We show here the projection of the galaxy distribution in galactic coordinates. From Ref. [28].

Today various telescopes – like the Hubble Space Telescope, the twin Keck telescopes in Hawaii and the European Southern Observatory telescopes in Chile – are exploring the most distant regions of the universe and discovering the first galaxies at large distances. The furthest galaxies observed so far are at redshifts of  $z \simeq 10$  (at a distance of 13.7 billion light years from Earth), whose light was emitted when the universe had only about 3% of its present age. Only a few galaxies are known at those redshifts, but there are at present various catalogs like the CfA and APM galaxy catalogs, and more recently the IRAS Point Source redshift Catalog, see Fig. 10, and Las Campanas redshift surveys, that study the spatial distribution of hundreds of thousands of galaxies up to distances of a billion light years, or  $z < 0.1$ , or the 2 degree Field Galaxy Redshift Survey (2dFGRS) and the Sloan Digital Sky Survey (SDSS), which reach  $z < 0.5$  and study millions of galaxies. These catalogs are telling us about the evolution of clusters and superclusters of galaxies in the universe, and already put constraints on the theory of structure formation. From these observations one can infer that most galaxies formed at redshifts of the order of 2 – 6; clusters of galaxies formed at redshifts of order 1, and superclusters are forming now. That is, cosmic structure formed from the bottom up: from galaxies to clusters to superclusters, and not the other way around. This fundamental difference is an indication of the type of matter that gave rise to structure.

We know from Big Bang nucleosynthesis that all the baryons in the universe cannot account for the observed amount of matter, so there must be some extra matter (dark since we don't see it) to account for its gravitational pull. Whether it is relativistic (hot) or non-relativistic (cold) could be inferred from observations: relativistic particles tend to diffuse from one concentration of matter to another, thus transferring energy among them and preventing the growth of structure on small scales. This is excluded by observations, so we conclude that most of the matter responsible for structure formation must be cold. How much there is is a matter of debate at the moment. Some recent analyses suggest that there is not enough cold dark matter to reach the critical density required to make the universe flat. If we want to

make sense of the present observations, we must conclude that some other form of energy permeates the universe. In order to resolve this issue, 2dFGRS and SDSS started taking data a few years ago. The first has already been completed, but the second one is still taking data up to redshifts  $z \simeq 5$  for quasars, over a large region of the sky. These important observations will help astronomers determine the nature of the dark matter and test the validity of the models of structure formation.

Before COBE discovered the anisotropies of the microwave background there were serious doubts whether gravity alone could be responsible for the formation of the structure we observe in the universe today. It seemed that a new force was required to do the job. Fortunately, the anisotropies were found with the right amplitude for structure to be accounted for by gravitational collapse of primordial inhomogeneities under the attraction of a large component of non-relativistic dark matter. Nowadays, the standard theory of structure formation is a cold dark matter model with a non vanishing cosmological constant in a spatially flat universe. Gravitational collapse amplifies the density contrast initially through linear growth and later on via non-linear collapse. In the process, overdense regions decouple from the Hubble expansion to become bound systems, which start attracting each other to form larger bound structures. In fact, the largest structures, superclusters, have not yet gone non-linear.

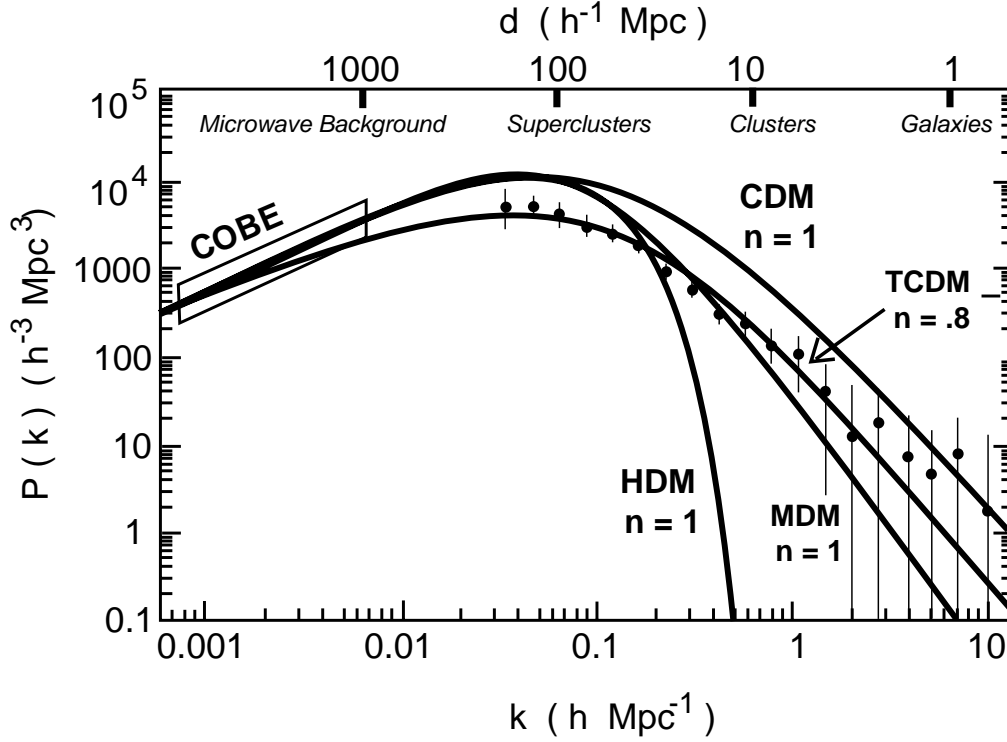
The primordial spectrum (100) is reprocessed by gravitational instability after the universe becomes matter dominated and inhomogeneities can grow. Linear perturbation theory shows that the growing mode<sup>5</sup> of small density contrasts go like  $\delta(a) \propto a$ , in the Einstein-de Sitter limit [25, 26]. There are slight deviations for  $a \gg a_{\text{eq}}$ , if  $\Omega_M \neq 1$  or  $\Omega_\Lambda \neq 0$ , but we will not be concerned with them here. The important observation is that, since the density contrast at last scattering is of order  $\delta \sim 10^{-5}$ , and the scale factor has grown since then only a factor  $z_{\text{dec}} \sim 10^3$ , one would expect a density contrast today of order  $\delta_0 \sim 10^{-2}$ . Instead, we observe structures like galaxies, where  $\delta \sim 10^6$ . So how can this be possible? The microwave background shows anisotropies due to fluctuations in the baryonic matter component only (to which photons couple, electromagnetically). If there is an additional matter component that only couples through very weak interactions, fluctuations in that component could grow as soon as it decoupled from the plasma, well before photons decoupled from baryons. The reason why baryonic inhomogeneities cannot grow is because of photon pressure: as baryons collapse towards denser regions, radiation pressure eventually halts the contraction and sets up acoustic oscillations in the plasma that prevent the growth of perturbations, until photon decoupling. On the other hand, a weakly interacting cold dark matter component could start gravitational collapse much earlier, even before matter-radiation equality, and thus reach the density contrast amplitudes observed today. The resolution of this mismatch is one of the strongest arguments for the existence of a weakly interacting cold dark matter component of the universe.

How much dark matter there is in the universe can be deduced from the actual power spectrum (the Fourier transform of the two-point correlation function of density perturbations) of the observed large scale structure. One can decompose the density contrast in Fourier components, see Eq. (99). This is very convenient since in linear perturbation theory individual Fourier components evolve independently. A comoving wavenumber  $k$  is said to “enter the horizon” when  $k = d_H^{-1}(a) = aH(a)$ . If a certain perturbation, of wavelength  $\lambda = k^{-1} < d_H(a_{\text{eq}})$ , enters the horizon before matter-radiation equality, the fast radiation-driven expansion prevents dark-matter perturbations from collapsing. Since light can only cross regions that are smaller than the horizon, the suppression of growth due to radiation is restricted to scales smaller than the horizon, while large-scale perturbations remain unaffected. This is the reason why the horizon size at equality, Eq. (95), sets an important scale for structure growth,

$$k_{\text{eq}} = d_H^{-1}(a_{\text{eq}}) \simeq 0.083 (\Omega_M h) h \text{ Mpc}^{-1}. \quad (101)$$

The suppression factor can be easily computed as  $f_{\text{sup}} = (a_{\text{enter}}/a_{\text{eq}})^2 = (k_{\text{eq}}/k)^2$ . In other words, the

<sup>5</sup>The decaying mode goes like  $\delta(t) \sim t^{-1}$ , for all  $\omega$ .



**Fig. 11:** The power spectrum for cold dark matter (CDM), tilted cold dark matter (TCDM), hot dark matter (HDM), and mixed hot plus cold dark matter (MDM), normalized to COBE, for large-scale structure formation. From Ref. [29].

processed power spectrum  $P(k)$  will have the form:

$$P(k) \propto \begin{cases} k, & k \ll k_{\text{eq}} \\ k^{-3}, & k \gg k_{\text{eq}} \end{cases} \quad (102)$$

This is precisely the shape that large-scale galaxy catalogs are bound to test in the near future, see Fig. 11. Furthermore, since relativistic Hot Dark Matter (HDM) transfer energy between clumps of matter, they will wipe out small scale perturbations, and this should be seen as a distinctive signature in the matter power spectra of future galaxy catalogs. On the other hand, non-relativistic Cold Dark Matter (CDM) allow structure to form on *all* scales via gravitational collapse. The dark matter will then pull in the baryons, which will later shine and thus allow us to see the galaxies.

Naturally, when baryons start to collapse onto dark matter potential wells, they will convert a large fraction of their potential energy into kinetic energy of protons and electrons, ionizing the medium. As a consequence, we expect to see a large fraction of those baryons constituting a hot ionized gas surrounding large clusters of galaxies. This is indeed what is observed, and confirms the general picture of structure formation.

### 3 Determination of Cosmological Parameters

In this Section, I will restrict myself to those recent measurements of the cosmological parameters by means of standard cosmological techniques, together with a few instances of new results from recently applied techniques. We will see that a large host of observations are determining the cosmological parameters with some reliability of the order of 10%. However, the majority of these measurements are dominated by large systematic errors. Most of the recent work in observational cosmology has been

the search for virtually systematic-free observables, like those obtained from the microwave background anisotropies, and discussed in Section 4.4. I will devote, however, this Section to the more ‘classical’ measurements of the following cosmological parameters: The rate of expansion  $H_0$ ; the matter content  $\Omega_M$ ; the cosmological constant  $\Omega_\Lambda$ ; the spatial curvature  $\Omega_K$ , and the age of the universe  $t_0$ .

### 3.1 The rate of expansion $H_0$

Over most of last century the value of  $H_0$  has been a constant source of disagreement [30]. Around 1929, Hubble measured the rate of expansion to be  $H_0 = 500 \text{ km s}^{-1}\text{Mpc}^{-1}$ , which implied an age of the universe of order  $t_0 \sim 2 \text{ Gyr}$ , in clear conflict with geology. Hubble’s data was based on Cepheid standard candles that were incorrectly calibrated with those in the Large Magellanic Cloud. Later on, in 1954 Baade recalibrated the Cepheid distance and obtained a lower value,  $H_0 = 250 \text{ km s}^{-1}\text{Mpc}^{-1}$ , still in conflict with ratios of certain unstable isotopes. Finally, in 1958 Sandage realized that the brightest stars in galaxies were ionized HII regions, and the Hubble rate dropped down to  $H_0 = 60 \text{ km s}^{-1}\text{Mpc}^{-1}$ , still with large (factor of two) systematic errors. Fortunately, in the past 15 years there has been significant progress towards the determination of  $H_0$ , with systematic errors approaching the 10% level. These improvements come from two directions. First, technological, through the replacement of photographic plates (almost exclusively the source of data from the 1920s to 1980s) with charged couple devices (CCDs), i.e. solid state detectors with excellent flux sensitivity per pixel, which were previously used successfully in particle physics detectors. Second, by the refinement of existing methods for measuring extragalactic distances (e.g. parallax, Cepheids, supernovae, etc.). Finally, with the development of completely new methods to determine  $H_0$ , which fall into totally independent and very broad categories: a) Gravitational lensing; b) Sunyaev-Zel’dovich effect; c) Extragalactic distance scale, mainly Cepheid variability and type Ia Supernovae; d) Microwave background anisotropies. I will review here the first three, and leave the last method for Section 4.4, since it involves knowledge about the primordial spectrum of inhomogeneities.

#### 3.1.1 Gravitational lensing

Imagine a quasi-stellar object (QSO) at large redshift ( $z \gg 1$ ) whose light is lensed by an intervening galaxy at redshift  $z \sim 1$  and arrives to an observer at  $z = 0$ . There will be at least two different images of the same background *variable* point source. The arrival times of photons from two different gravitationally lensed images of the quasar depend on the different path lengths and the gravitational potential traversed. Therefore, a measurement of the time delay and the angular separation of the different images of a variable quasar can be used to determine  $H_0$  with great accuracy. This method, proposed in 1964 by Refsdal [31], offers tremendous potential because it can be applied at great distances and it is based on very solid physical principles [32].

Unfortunately, there are very few systems with both a favourable geometry (i.e. a known mass distribution of the intervening galaxy) and a variable background source with a measurable time delay. That is the reason why it has taken so much time since the original proposal for the first results to come out. Fortunately, there are now very powerful telescopes that can be used for these purposes. The best candidate to-date is the QSO 0957 + 561, observed with the 10m Keck telescope, for which there is a model of the lensing mass distribution that is consistent with the measured velocity dispersion. Assuming a flat space with  $\Omega_M = 0.25$ , one can determine [33]

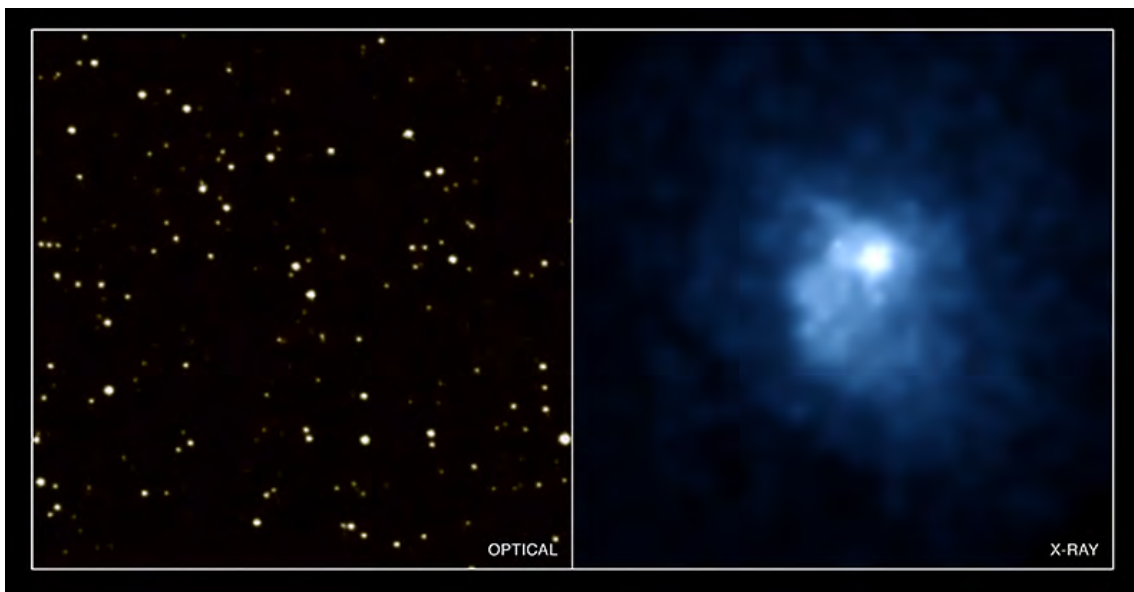
$$H_0 = 72 \pm 7 (1\sigma \text{ statistical}) \pm 15\% (\text{systematic}) \text{ km s}^{-1}\text{Mpc}^{-1}. \quad (103)$$

The main source of systematic error is the degeneracy between the mass distribution of the lens and the value of  $H_0$ . Knowledge of the velocity dispersion within the lens as a function of position helps constrain the mass distribution, but those measurements are very difficult and, in the case of lensing by a cluster of galaxies, the dark matter distribution in those systems is usually unknown, associated with a

complicated cluster potential. Nevertheless, the method is just starting to give promising results and, in the near future, with the recent discovery of several systems with optimum properties, the prospects for measuring  $H_0$  and lowering its uncertainty with this technique are excellent.

### 3.1.2 Sunyaev-Zel'dovich effect

As discussed in the previous Section, the gravitational collapse of baryons onto the potential wells generated by dark matter gave rise to the reionization of the plasma, generating an X-ray halo around rich clusters of galaxies, see Fig. 12. The inverse-Compton scattering of microwave background photons off the hot electrons in the X-ray gas results in a measurable distortion of the blackbody spectrum of the microwave background, known as the Sunyaev-Zel'dovich (SZ) effect. Since photons acquire extra energy from the X-ray electrons, we expect a shift towards higher frequencies of the spectrum,  $(\Delta\nu/\nu) \simeq (k_B T_{\text{gas}}/m_e c^2) \sim 10^{-2}$ . This corresponds to a *decrement* of the microwave background temperature at low frequencies (Rayleigh-Jeans region) and an increment at high frequencies, see Ref. [34].



**Fig. 12:** The 3C438 cluster of galaxies, seen here in an optical image (left) and an X-ray image (right), taken by Chandra X-ray Observatory. It is clear that the gas in the center of the cluster is very hot and has no optical counterpart with any particular galaxy. From Ref. [35].

Measuring the *spatial* distribution of the SZ effect (3 K spectrum), together with a high resolution X-ray map ( $10^8$  K spectrum) of the cluster, one can determine the density and temperature distribution of the hot gas. Since the X-ray flux is distance-dependent ( $\mathcal{F} = \mathcal{L}/4\pi d_L^2$ ), while the SZ decrement is not (because the energy of the CMB photons increases as we go back in redshift,  $\nu = \nu_0(1+z)$ , and exactly compensates the redshift in energy of the photons that reach us), one can determine from there the distance to the cluster, and thus the Hubble rate  $H_0$ .

The advantages of this method are that it can be applied to large distances and it is based on clear physical principles. The main systematics come from possible clumpiness of the gas (which would reduce  $H_0$ ), projection effects (if the clusters are prolate,  $H_0$  could be larger), the assumption of hydrostatic equilibrium of the X-ray gas, details of models for the gas and electron densities, and possible contaminations from point sources. Present measurements give the value [34]

$$H_0 = 60 \pm 10 (1\sigma \text{ statistical}) \pm 20\% (\text{systematic}) \text{ km s}^{-1}\text{Mpc}^{-1}, \quad (104)$$

compatible with other determinations. A great advantage of this completely new and independent method is that nowadays more and more clusters are observed in the X-ray, and soon we will have high-resolution 2D maps of the SZ decrement from several balloon flights, as well as from future microwave background satellites, together with precise X-ray maps and spectra from the Chandra X-ray observatory recently launched by NASA, as well as from the European X-ray satellite XMM launched a few months ago by ESA, which will deliver orders of magnitude better resolution than the existing Einstein X-ray satellite.

### 3.1.3 Cepheid variability

Cepheids are low-mass variable stars with a period-luminosity relation based on the helium ionization cycles inside the star, as it contracts and expands. This time variability can be measured, and the star's absolute luminosity determined from the calibrated relationship. From the observed flux one can then deduce the luminosity distance, see Eq. (28), and thus the Hubble rate  $H_0$ . The Hubble Space Telescope (HST) was launched by NASA in 1990 (and repaired in 1993) with the specific project of calibrating the extragalactic distance scale and thus determining the Hubble rate with 10% accuracy. The most recent results from HST are the following [36]

$$H_0 = 71 \pm 4 \text{ (random)} \pm 7 \text{ (systematic)} \text{ km s}^{-1}\text{Mpc}^{-1}. \quad (105)$$

The main source of systematic error is the distance to the Large Magellanic Cloud, which provides the fiducial comparison for Cepheids in more distant galaxies. Other systematic uncertainties that affect the value of  $H_0$  are the internal extinction correction method used, a possible metallicity dependence of the Cepheid period-luminosity relation and cluster population incompleteness bias, for a set of 21 galaxies within 25 Mpc, and 23 clusters within  $z \lesssim 0.03$ .

With better telescopes already taking data, like the Very Large Telescope (VLT) interferometer of the European Southern Observatory (ESO) in the Chilean Atacama desert, with 8 synchronized telescopes, and others coming up soon, like the Next Generation Space Telescope (NGST) proposed by NASA for 2008, and the Gran TeCan of the European Northern Observatory in the Canary Islands, for 2010, it is expected that much better resolution and therefore accuracy can be obtained for the determination of  $H_0$ .

## 3.2 Dark Matter

In the 1920s Hubble realized that the so called nebulae were actually distant galaxies very similar to our own. Soon afterwards, in 1933, Zwicky found dynamical evidence that there is possibly ten to a hundred times more mass in the Coma cluster than contributed by the luminous matter in galaxies [37]. However, it was not until the 1970s that the existence of dark matter began to be taken more seriously. At that time there was evidence that rotation curves of galaxies did not fall off with radius and that the dynamical mass was increasing with scale from that of individual galaxies up to clusters of galaxies. Since then, new possible extra sources to the matter content of the universe have been accumulating:

$$\Omega_M = \Omega_{B, \text{lum}} \quad (\text{stars in galaxies}) \quad (106)$$

$$+ \Omega_{B, \text{dark}} \quad (\text{MACHOs?}) \quad (107)$$

$$+ \Omega_{CDM} \quad (\text{weakly interacting : axion, neutralino?}) \quad (108)$$

$$+ \Omega_{HDM} \quad (\text{massive neutrinos?}) \quad (109)$$

The empirical route to the determination of  $\Omega_M$  is nowadays one of the most diversified of all cosmological parameters. The matter content of the universe can be deduced from the mass-to-light ratio of various objects in the universe; from the rotation curves of galaxies; from microlensing and the direct search of Massive Compact Halo Objects (MACHOs); from the cluster velocity dispersion with the use of the Virial theorem; from the baryon fraction in the X-ray gas of clusters; from weak gravitational

lensing; from the observed matter distribution of the universe via its power spectrum; from the cluster abundance and its evolution; from direct detection of massive neutrinos at SuperKamiokande; from direct detection of Weakly Interacting Massive Particles (WIMPs) at CDMS, DAMA or UKDMC, and finally from microwave background anisotropies. I will review here just a few of them.

### 3.2.1 *Rotation curves of spiral galaxies*

The flat rotation curves of spiral galaxies provide the most direct evidence for the existence of large amounts of dark matter. Spiral galaxies consist of a central bulge and a very thin disk, stabilized against gravitational collapse by angular momentum conservation, and surrounded by an approximately spherical halo of dark matter. One can measure the orbital velocities of objects orbiting around the disk as a function of radius from the Doppler shifts of their spectral lines.

The rotation curve of the Andromeda galaxy was first measured by Babcock in 1938, from the stars in the disk. Later it became possible to measure galactic rotation curves far out into the disk, and a trend was found [39, 40]. The orbital velocity rose linearly from the center outward until it reached a typical value of 200 km/s, and then remained flat out to the largest measured radii. This was completely unexpected since the observed surface luminosity of the disk falls off exponentially with radius [39],  $I(r) = I_0 \exp(-r/r_D)$ . Therefore, one would expect that most of the galactic mass is concentrated within a few disk lengths  $r_D$ , such that the rotation velocity is determined as in a Keplerian orbit,  $v_{\text{rot}} = (GM/r)^{1/2} \propto r^{-1/2}$ . No such behaviour is observed. In fact, the most convincing observations come from radio emission (from the 21 cm line) of neutral hydrogen in the disk, which has been measured to much larger galactic radii than optical tracers. A typical case is that of the spiral galaxy NGC 6503, where  $r_D = 1.73$  kpc, while the furthest measured hydrogen line is at  $r = 22.22$  kpc, about 13 disk lengths away. Nowadays, thousands of galactic rotation curves are known, see Fig. 14, and all suggest the existence of about ten times more mass in the halos of spiral galaxies than in the stars of the disk. The connection with dark matter halos was emphasized in Ref. [41]. Recent numerical simulations of galaxy formation in a CDM cosmology [42] suggest that galaxies probably formed by the infall of material in an overdense region of the universe that had decoupled from the overall expansion.

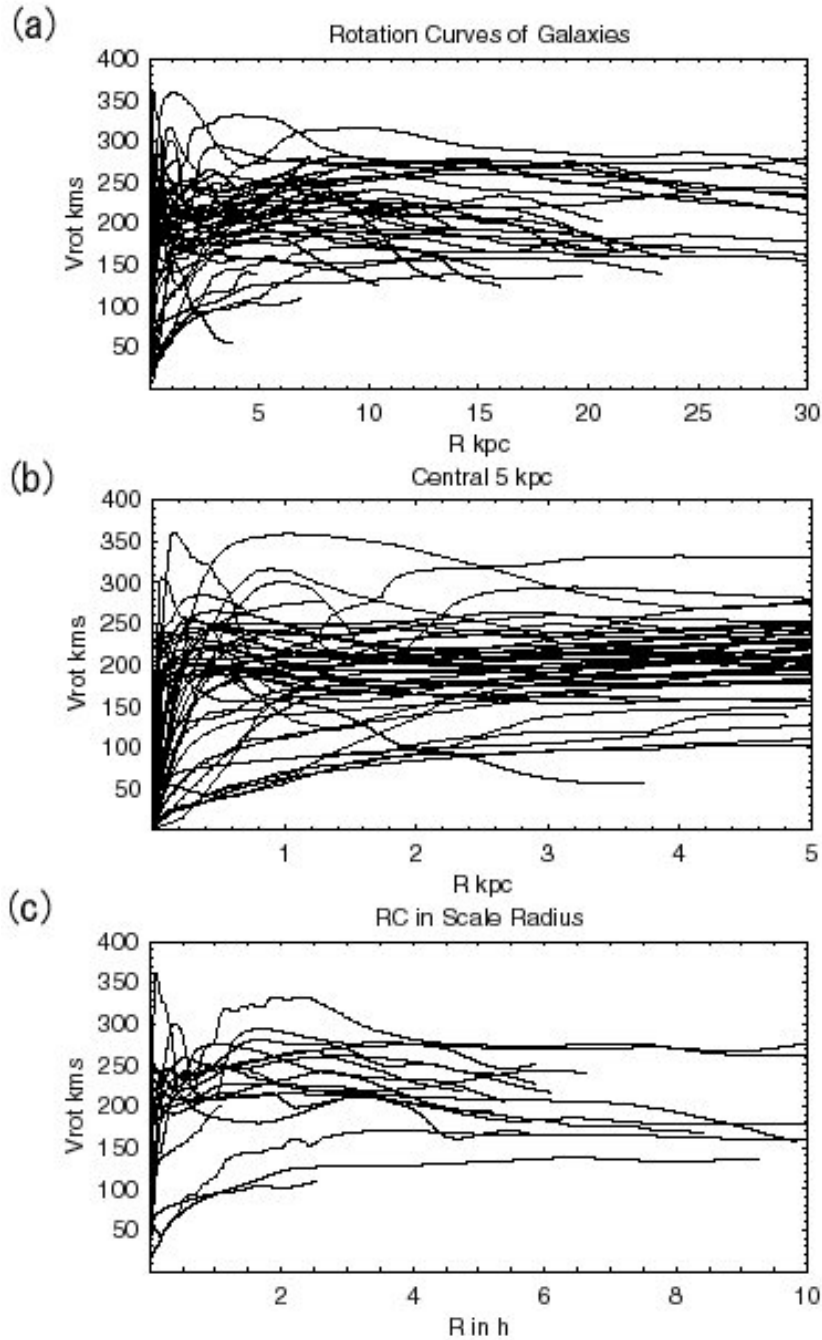
The dark matter is supposed to undergo violent relaxation and create a virialized system, i.e. in hydrostatic equilibrium. This picture has led to a simple model of dark-matter halos as isothermal spheres, with density profile  $\rho(r) = \rho_c / (r_c^2 + r^2)$ , where  $r_c$  is a core radius and  $\rho_c = v_\infty^2 / 4\pi G$ , with  $v_\infty$  equal to the plateau value of the flat rotation curve. This model is consistent with the universal rotation curves seen in Fig. 6. At large radii the dark matter distribution leads to a flat rotation curve. The question is for how long. In dense galaxy clusters one expects the galactic halos to overlap and form a continuum, and therefore the rotation curves should remain flat from one galaxy to another. However, in field galaxies, far from clusters, one can study the rotation velocities of substructures (like satellite dwarf galaxies) around a given galaxy, and determine whether they fall off at sufficiently large distances according to Kepler's law, as one would expect, once the edges of the dark matter halo have been reached. These observations are rather difficult because of uncertainties in distinguishing between true satellites and interlopers. Recently, a group from the Sloan Digital Sky Survey Collaboration claim that they have seen the edges of the dark matter halos around field galaxies by confirming the fall-off at large distances of their rotation curves [43]. These results, if corroborated by further analysis, would constitute a tremendous support to the idea of dark matter as a fluid surrounding galaxies and clusters, while at the same time eliminates the need for modifications of Newtonian or even Einsteinian gravity at the scales of galaxies, to account for the flat rotation curves.

That's fine, but how much dark matter is there at the galactic scale? Adding up all the matter in galactic halos up to a maximum radii, one finds

$$\Omega_{\text{halo}} \simeq 10 \Omega_{\text{lum}} \geq 0.03 - 0.05. \quad (110)$$

Of course, it would be extraordinary if we could confirm, through direct detection, the existence of dark





**Fig. 13:** The rotation curves of several hundred galaxies. Upper panel: As a function of their radii in kpc. Middle panel: The central 5 kpc. Lower panel: As a function of scale radius.

matter in our own galaxy. For that purpose, one should measure its rotation curve, which is much more difficult because of obscuration by dust in the disk, as well as problems with the determination of reliable galactocentric distances for the tracers. Nevertheless, the rotation curve of the Milky Way has been measured and conforms to the usual picture, with a plateau value of the rotation velocity of 220 km/s. For dark matter searches, the crucial quantity is the dark matter density in the solar neighbourhood, which turns out to be (within a factor of two uncertainty depending on the halo model)  $\rho_{\text{DM}} = 0.3 \text{ GeV/cm}^3$ . We will come back to direct searched of dark matter in a later subsection.

### 3.2.2 Baryon fraction in clusters

Since large clusters of galaxies form through gravitational collapse, they scoop up mass over a large volume of space, and therefore the ratio of baryons over the total matter in the cluster should be representative of the entire universe, at least within a 20% systematic error. Since the 1960s, when X-ray telescopes became available, it is known that galaxy clusters are the most powerful X-ray sources in the sky [44]. The emission extends over the whole cluster and reveals the existence of a hot plasma with temperature  $T \sim 10^7 - 10^8$  K, where X-rays are produced by electron bremsstrahlung. Assuming the gas to be in hydrostatic equilibrium and applying the virial theorem one can estimate the total mass in the cluster, giving general agreement (within a factor of 2) with the virial mass estimates. From these estimates one can calculate the baryon fraction of clusters

$$f_B h^{3/2} = 0.08 \quad \Rightarrow \quad \frac{\Omega_B}{\Omega_M} \approx 0.14, \quad \text{for } h = 0.70. \quad (111)$$

Since  $\Omega_{\text{lum}} \simeq 0.002 - 0.006$ , the previous expression suggests that clusters contain far more baryonic matter in the form of hot gas than in the form of stars in galaxies. Assuming this fraction to be representative of the entire universe, and using the Big Bang nucleosynthesis value of  $\Omega_B = 0.04 \pm 0.01$ , for  $h = 0.7$ , we find

$$\Omega_M = 0.3 \pm 0.1 \text{ (statistical)} \pm 20\% \text{ (systematic)}. \quad (112)$$

This value is consistent with previous determinations of  $\Omega_M$ . If some baryons are ejected from the cluster during gravitational collapse, or some are actually bound in nonluminous objects like planets, then the actual value of  $\Omega_M$  is smaller than this estimate.

### 3.2.3 Weak gravitational lensing

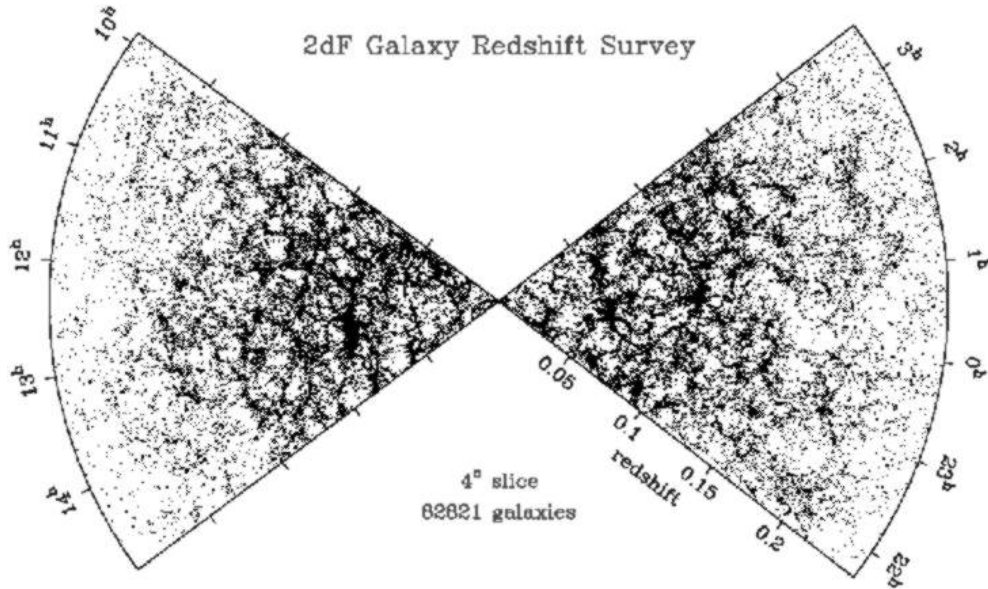
Since the mid 1980s, deep surveys with powerful telescopes have observed huge arc-like features in galaxy clusters. The spectroscopic analysis showed that the cluster and the giant arcs were at very different redshifts. The usual interpretation is that the arc is the image of a distant background galaxy which is in the same line of sight as the cluster so that it appears distorted and magnified by the gravitational lens effect: the giant arcs are essentially partial Einstein rings. From a systematic study of the cluster mass distribution one can reconstruct the shear field responsible for the gravitational distortion [45]. This analysis shows that there are large amounts of dark matter in the clusters, in rough agreement with the virial mass estimates, although the lensing masses tend to be systematically larger. At present, the estimates indicate  $\Omega_M = 0.2 - 0.3$  on scales  $\lesssim 6 h^{-1}$  Mpc.

### 3.2.4 Large scale structure formation and the matter power spectrum

Although the isotropic microwave background indicates that the universe in the *past* was extraordinarily homogeneous, we know that the universe *today* is far from homogeneous: we observe galaxies, clusters and superclusters on large scales. These structures are expected to arise from very small primordial inhomogeneities that grow in time via gravitational instability, and that may have originated from tiny ripples in the metric, as matter fell into their troughs. Those ripples must have left some trace as temperature anisotropies in the microwave background, and indeed such anisotropies were finally discovered by the COBE satellite in 1992. However, not all kinds of matter and/or evolution of the universe can give rise to the structure we observe today. If we define the density contrast as

$$\delta(\vec{x}, a) \equiv \frac{\rho(\vec{x}, a) - \bar{\rho}(a)}{\bar{\rho}(a)} = \int d^3\vec{k} \delta_k(a) e^{i\vec{k}\cdot\vec{x}}, \quad (113)$$

where  $\bar{\rho}(a) = \rho_0 a^{-3}$  is the average cosmic density, we need a theory that will grow a density contrast with amplitude  $\delta \sim 10^{-5}$  at the last scattering surface ( $z = 1100$ ) up to density contrasts of the order of



**Fig. 14:** The 2 degree Field Galaxy Redshift Survey contains some 250,000 galaxies, covering a large fraction of the sky up to redshifts of  $z \leq 0.25$ . From Ref. [46].

$\delta \sim 10^6$  for galaxies at redshifts  $z \ll 1$ , i.e. today. This is a *necessary* requirement for any consistent theory of structure formation.

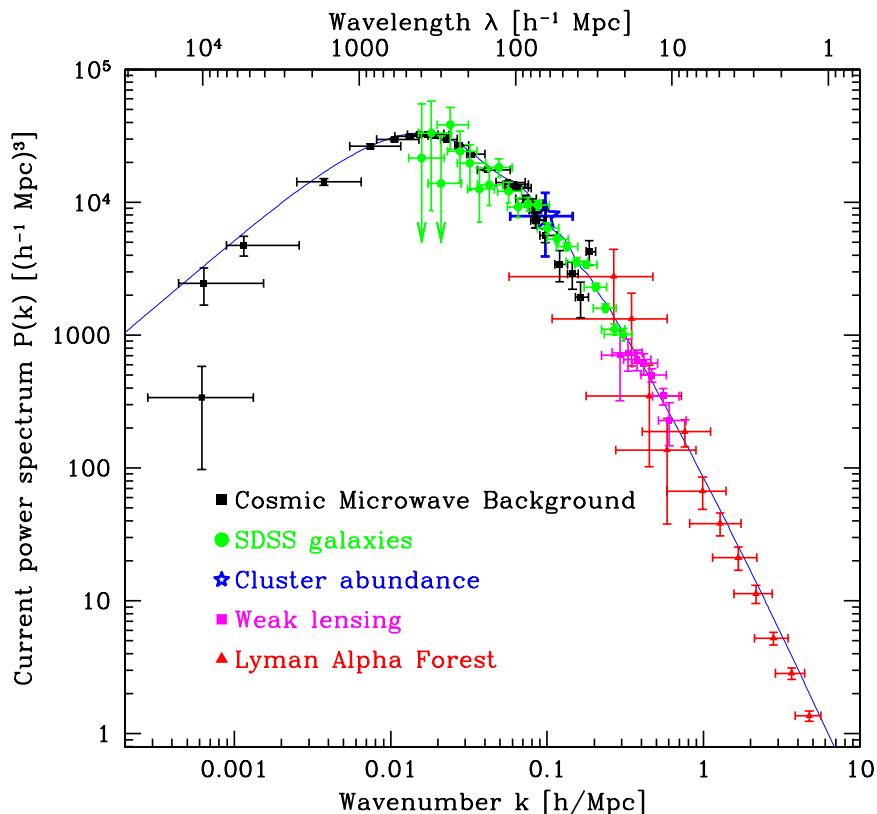
Furthermore, the anisotropies observed by the COBE satellite correspond to a small-amplitude scale-invariant primordial power spectrum of inhomogeneities

$$P(k) = \langle |\delta_k|^2 \rangle \propto k^n, \quad \text{with } n = 1, \quad (114)$$

These inhomogeneities are like waves in the space-time metric. When matter fell in the troughs of those waves, it created density perturbations that collapsed gravitationally to form galaxies and clusters of galaxies, with a spectrum that is also scale invariant. Such a type of spectrum was proposed in the early 1970s by Edward R. Harrison, and independently by the Russian cosmologist Yakov B. Zel'dovich [27], to explain the distribution of galaxies and clusters of galaxies on very large scales in our observable universe, see Fig. 14.

Since the primordial spectrum is very approximately represented by a scale-invariant *Gaussian random field*, the best way to present the results of structure formation is by working with the 2-point correlation function in Fourier space, the so-called *power spectrum*. If the reprocessed spectrum of inhomogeneities remains Gaussian, the power spectrum is all we need to describe the galaxy distribution. Non-Gaussian effects are expected to arise from the non-linear gravitational collapse of structure, and may be important at small scales. The power spectrum measures the degree of inhomogeneity in the mass distribution on different scales, see Fig. 15. It depends upon a few basic ingredients: a) the primordial spectrum of inhomogeneities, whether they are Gaussian or non-Gaussian, whether *adiabatic* (perturbations in the energy density) or *isocurvature* (perturbations in the entropy density), whether the primordial spectrum has *tilt* (deviations from scale-invariance), etc.; b) the recent creation of inhomogeneities, whether *cosmic strings* or some other topological defect from an early phase transition are responsible for the formation of structure today; and c) the cosmic evolution of the inhomogeneity, whether the universe has been dominated by cold or hot dark matter or by a cosmological constant since the beginning of structure formation, and also depending on the rate of expansion of the universe.

The working tools used for the comparison between the observed power spectrum and the predicted one are very precise N-body numerical simulations and theoretical models that predict the *shape*

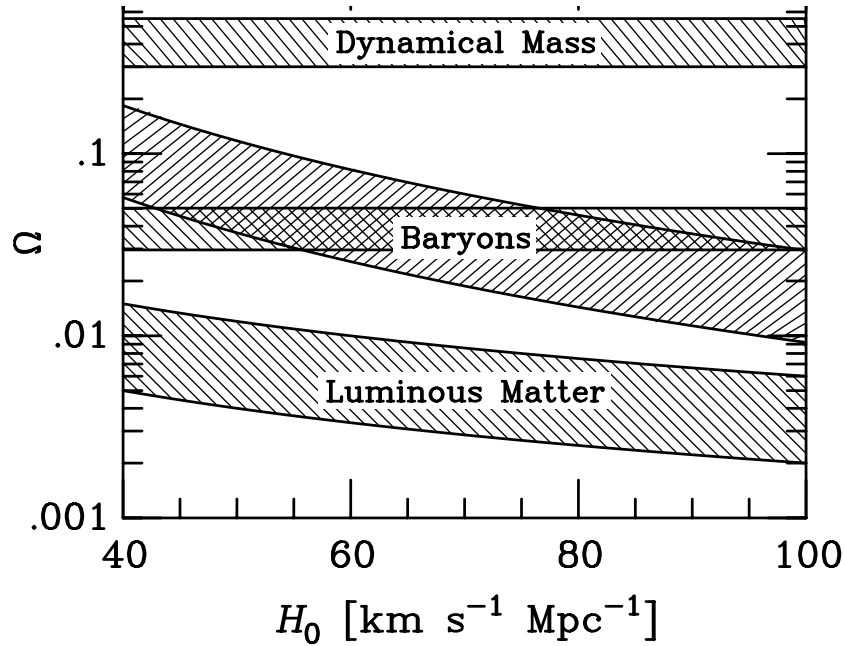


**Fig. 15:** The measured power spectrum  $P(k)$  as a function of wavenumber  $k$ . From observations of the Sloan Digital Sky Survey, CMB anisotropies, cluster abundance, gravitational lensing and Lyman- $\alpha$  forest. From Ref. [47].

but not the *amplitude* of the present power spectrum. Even though a large amount of work has gone into those analyses, we still have large uncertainties about the nature and amount of matter necessary for structure formation. A model that has become a working paradigm is a flat cold dark matter model with a cosmological constant and  $\Omega_M \sim 0.3$ . This model is now being confronted with the recent very precise measurements from 2dFGRS [46] and SDSS [47].

### 3.2.5 The new redshift catalogs, 2dF and Sloan Digital Sky Survey

Our view of the large-scale distribution of luminous objects in the universe has changed dramatically during the last 25 years: from the simple pre-1975 picture of a distribution of field and cluster galaxies, to the discovery of the first single superstructures and voids, to the most recent results showing an almost regular web-like network of interconnected clusters, filaments and walls, separating huge nearly empty volumes. The increased efficiency of redshift surveys, made possible by the development of spectrographs and – specially in the last decade – by an enormous increase in multiplexing gain (i.e. the ability to collect spectra of several galaxies at once, thanks to fibre-optic spectrographs), has allowed us not only to do *cartography* of the nearby universe, but also to statistically characterize some of its properties. At the same time, advances in theoretical modeling of the development of structure, with large high-resolution gravitational simulations coupled to a deeper yet limited understanding of how to form galaxies within the dark matter halos, have provided a more realistic connection of the models to the observable quantities. Despite the large uncertainties that still exist, this has transformed the study of cosmology and large-scale structure into a truly quantitative science, where theory and observations can progress together.



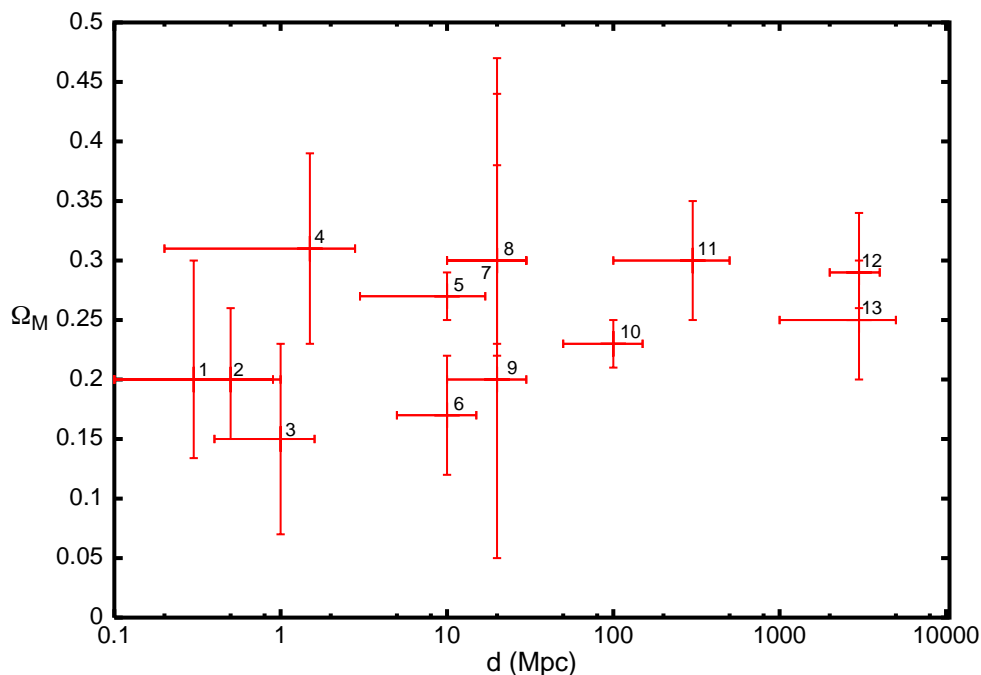
**Fig. 16:** The observed cosmic matter components as functions of the Hubble expansion parameter. The luminous matter component is given by  $0.002 \leq \Omega_{\text{lum}} \leq 0.006$ ; the galactic halo component is the horizontal band,  $0.03 \leq \Omega_{\text{halo}} \leq 0.05$ , crossing the baryonic component from BBN,  $\Omega_B h^2 = 0.0244 \pm 0.0024$ ; and the dynamical mass component from large scale structure analysis is given by  $\Omega_M = 0.3 \pm 0.1$ . Note that in the range  $H_0 = 71 \pm 3$  km/s/Mpc, there are *three* dark matter problems, see the text. From Ref. [48].

### 3.2.6 Summary of the matter content

We can summarize the present situation with Fig. 16, for  $\Omega_M$  as a function of  $H_0$ . There are four bands, the luminous matter  $\Omega_{\text{lum}}$ ; the baryon content  $\Omega_B$ , from BBN; the galactic halo component  $\Omega_{\text{halo}}$ , and the dynamical mass from clusters,  $\Omega_M$ . From this figure it is clear that there are in fact *three* dark matter problems: The first one is where are 90% of the baryons? Between the fraction predicted by BBN and that seen in stars and diffuse gas there is a huge fraction which is in the form of dark baryons. They could be in small clumps of hydrogen that have not started thermonuclear reactions and perhaps constitute the dark matter of spiral galaxies' halos. Note that although  $\Omega_B$  and  $\Omega_{\text{halo}}$  coincide at  $H_0 \simeq 70$  km/s/Mpc, this could be just a coincidence. The second problem is what constitutes 90% of matter, from BBN baryons to the mass inferred from cluster dynamics? This is the standard dark matter problem and could be solved in the future by direct detection of a weakly interacting massive particle in the laboratory. And finally, since we know from observations of the CMB that the universe is flat, the rest, up to  $\Omega_0 = 1$ , must be a diffuse vacuum energy, which affects the very large scales and late times, and seems to be responsible for the present acceleration of the universe, see Section 3. Nowadays, multiple observations seem to converge towards a common determination of  $\Omega_M = 0.25 \pm 0.08$  (95% c.l.), see Fig. 17.

### 3.2.7 Massive neutrinos

One of the 'usual suspects' when addressing the problem of dark matter are neutrinos. They are the only candidates known to exist. If neutrinos have a mass, could they constitute the missing matter? We know from the Big Bang theory, see Section 2.6.5, that there is a cosmic neutrino background at a temperature of approximately 2K. This allows one to compute the present number density in the form of neutrinos, which turns out to be, for massless neutrinos,  $n_\nu(T_\nu) = \frac{3}{11} n_\gamma(T_\gamma) = 112 \text{ cm}^{-3}$ , per species of neutrino. Since neutrinos have mass, see Fig. 18, the cosmic energy density in massive neutrinos



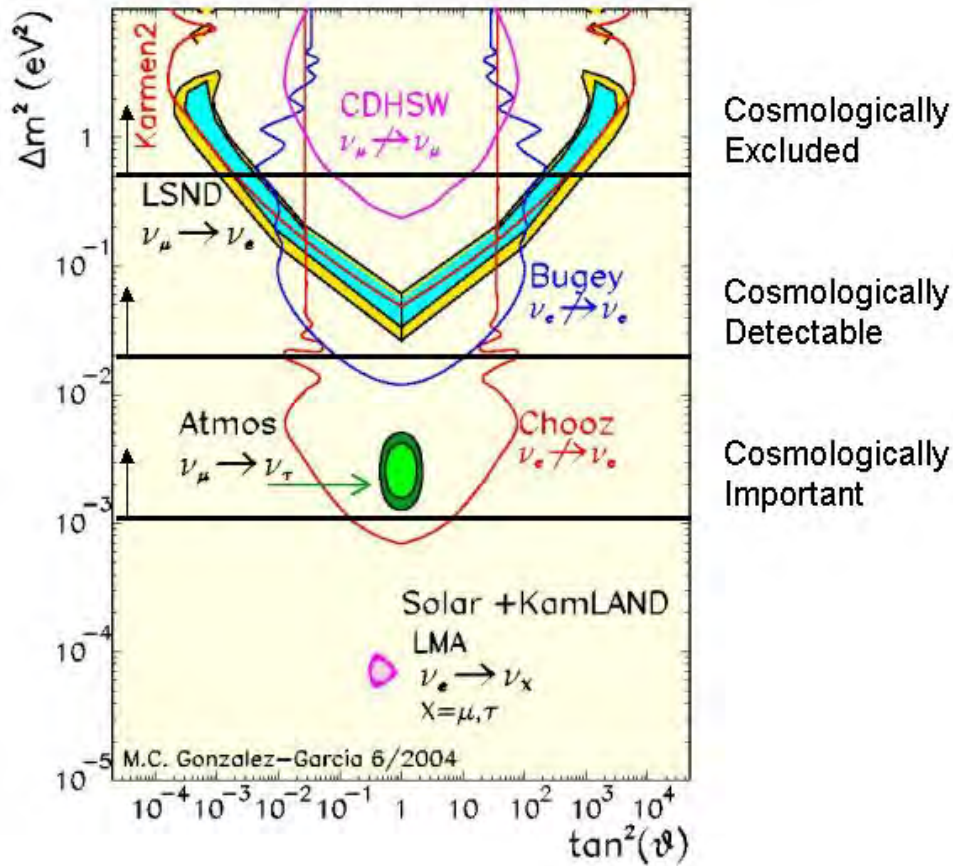
**Fig. 17:** Different determinations of  $\Omega_M$  as a function of distance, from various sources: 1. peculiar velocities; 2. weak gravitational lensing; 3. shear autocorrelation function; 4. local group of galaxies; 5. baryon mass fraction; 6. cluster mass function; 7. virgocentric flow; 8. mean relative velocities; 9. redshift space distortions; 10. mass power spectrum; 11. integrated Sachs-Wolfe effect; 12. angular diameter distance: SNe; 13. cluster baryon fraction. While a few years ago the dispersion among observed values was huge and strongly dependent on scale, at present the observed value of the matter density parameter falls well within a narrow range,  $\Omega_M = 0.25 \pm 0.07$  (95% c.l.) and is essentially independent on scale, from 100 kpc to 5000 Mpc. Adapted from Ref. [49].

would be  $\rho_\nu = \sum n_\nu m_\nu = \frac{3}{11} n_\gamma \sum m_\nu$ , and therefore its contribution today,

$$\Omega_\nu h^2 = \frac{\sum m_\nu}{93.2 \text{ eV}}. \quad (115)$$

The discussion in the previous Sections suggest that  $\Omega_M \leq 0.4$ , and thus, for any of the three families of neutrinos,  $m_\nu \leq 40 \text{ eV}$ . Note that this limit improves by six orders of magnitude the present bound on the tau-neutrino mass [20]. Supposing that the missing mass in non-baryonic cold dark matter arises from a single particle dark matter (PDM) component, its contribution to the critical density is bounded by  $0.05 \leq \Omega_{\text{PDM}} h^2 \leq 0.4$ , see Fig. 17.

I will now go through the various logical arguments that exclude neutrinos as the *dominant* component of the missing dark matter in the universe. Is it possible that neutrinos with a mass  $4 \text{ eV} \leq m_\nu \leq 40 \text{ eV}$  be the non-baryonic PDM component? For instance, could massive neutrinos constitute the dark matter halos of galaxies? For neutrinos to be gravitationally bound to galaxies it is necessary that their velocity be less than the escape velocity  $v_{\text{esc}}$ , and thus their maximum momentum is  $p_{\text{max}} = m_\nu v_{\text{esc}}$ . How many neutrinos can be packed in the halo of a galaxy? Due to the Pauli exclusion principle, the maximum number density is given by that of a completely degenerate Fermi gas with momentum  $p_F = p_{\text{max}}$ , i.e.  $n_{\text{max}} = p_{\text{max}}^3 / 3\pi^2$ . Therefore, the maximum local density in dark matter neutrinos is  $\rho_{\text{max}} = n_{\text{max}} m_\nu = m_\nu^4 v_{\text{esc}}^3 / 3\pi^2$ , which must be greater than the typical halo density  $\rho_{\text{halo}} = 0.3 \text{ GeV cm}^{-3}$ . For a typical spiral galaxy, this constraint, known as the Tremaine-Gunn limit, gives  $m_\nu \geq 40 \text{ eV}$ , see Ref. [51]. However, this mass, even for a single species, say the tau-neutrino,



**Fig. 18:** The neutrino parameter space, mixing angle against  $\Delta m^2$ , including the results from the different solar and atmospheric neutrino oscillation experiments. Note the threshold of cosmologically important masses, cosmologically detectable neutrinos (by CMB and LSS observations), and cosmologically excluded range of masses. Adapted from Refs. [50] and [95].

gives a value for  $\Omega_\nu h^2 = 0.5$ , which is far too high for structure formation. Neutrinos of such a low mass would constitute a relativistic hot dark matter component, which would wash-out structure below the supercluster scale, against evidence from present observations, see Fig. 18. Furthermore, applying the same phase-space argument to the neutrinos as dark matter in the halo of dwarf galaxies gives  $m_\nu \geq 100$  eV, beyond closure density (115). We must conclude that the simple idea that light neutrinos could constitute the particle dark matter on all scales is ruled out. They could, however, still play a role as a sub-dominant hot dark matter component in a flat CDM model. In that case, a neutrino mass of order 1 eV is not cosmological excluded, see Fig. 18.

Another possibility is that neutrinos have a large mass, of order a few GeV. In that case, their number density at decoupling, see Section 2.5.1, is suppressed by a Boltzmann factor,  $\sim \exp(-m_\nu/T_{\text{dec}})$ . For masses  $m_\nu > T_{\text{dec}} \simeq 0.8$  MeV, the present energy density has to be computed as a solution of the corresponding Boltzmann equation. Apart from a logarithmic correction, one finds  $\Omega_\nu h^2 \simeq 0.1(10 \text{ GeV}/m_\nu)^2$  for Majorana neutrinos and slightly smaller for Dirac neutrinos. In either case, neutrinos could be the dark matter only if their mass was a few GeV. Laboratory limits for  $\nu_\tau$  of around 18 MeV [20], and much more stringent ones for  $\nu_\mu$  and  $\nu_e$ , exclude the known light neutrinos. However, there is always the possibility of a fourth unknown heavy and stable (perhaps sterile) neutrino. If it couples to the Z boson and has a mass below 45 GeV for Dirac neutrinos (39.5 GeV for Majorana neu-

trinos), then it is ruled out by measurements at LEP of the invisible width of the Z. There are two logical alternatives, either it is a sterile neutrino (it does not couple to the Z), or it does couple but has a larger mass. In the case of a Majorana neutrino (its own antiparticle), their abundance, for this mass range, is too small for being cosmologically relevant,  $\Omega_\nu h^2 \leq 0.005$ . If it were a Dirac neutrino there could be a lepton asymmetry, which may provide a higher abundance (similar to the case of baryogenesis). However, neutrinos scatter on nucleons via the weak axial-vector current (spin-dependent) interaction. For the small momentum transfers imparted by galactic WIMPs, such collisions are essentially coherent over an entire nucleus, leading to an enhancement of the effective cross section. The relatively large detection rate in this case allows one to exclude fourth-generation Dirac neutrinos for the galactic dark matter [52]. Anyway, it would be very implausible to have such a massive neutrino today, since it would have to be stable, with a life-time greater than the age of the universe, and there is no theoretical reason to expect a massive sterile neutrino that does not oscillate into the other neutrinos.

Of course, the definitive test to the possible contribution of neutrinos to the overall density of the universe would be to measure *directly* their mass in laboratory experiments. There are at present two types of experiments: neutrino oscillation experiments, which measure only *differences* in squared masses, and direct mass-searches experiments, like the tritium  $\beta$ -spectrum and the neutrinoless double- $\beta$  decay experiments, which measure directly the mass of the electron neutrino. The former experiments give a bound  $m_{\nu_e} \lesssim 2.3$  eV (95% c.l.) [53], while the latter claim [54] they have a positive evidence for a Majorana neutrino of mass  $m_\nu = 0.05 - 0.89$  eV (95% c.l.), although this result still awaits confirmation by other experiments. Neutrinos with such a mass could very well constitute the HDM component of the universe,  $\Omega_{\text{HDM}} \lesssim 0.15$ . The oscillation experiments give a range of possibilities for  $\Delta m_\nu^2 = 0.3 - 3$  eV<sup>2</sup> from LSND (not yet confirmed by MiniBooNE), to the atmospheric neutrino oscillations from SuperKamiokande ( $\Delta m_\nu^2 \simeq 2.2 \pm 0.5 \times 10^{-3}$  eV<sup>2</sup>,  $\tan^2 \theta = 1.0 \pm 0.3$ ) and the solar neutrino oscillations from KamLAND and the Sudbury Neutrino Observatory ( $\Delta m_\nu^2 \simeq 8.2 \pm 0.3 \times 10^{-5}$  eV<sup>2</sup>,  $\tan^2 \theta = 0.39 \pm 0.05$ ), see Ref. [50]. Only the first two possibilities would be cosmologically relevant, see Fig. 18. Thanks to recent observations by WMAP, SDSS and Planck, we can put stringent limits on the absolute scale of neutrino masses, see below (Section 3.4).

### 3.2.8 Weakly Interacting Massive Particles

Unless we drastically change the theory of gravity on large scales, baryons cannot make up the bulk of the dark matter. Massive neutrinos are the only alternative among the known particles, but they are essentially ruled out as a universal dark matter candidate, even if they may play a subdominant role as a hot dark matter component. There remains the mystery of what is the physical nature of the dominant cold dark matter component. Something like a heavy stable neutrino, a generic Weakly Interacting Massive Particle (WIMP), could be a reasonable candidate because its present abundance could fall within the expected range,

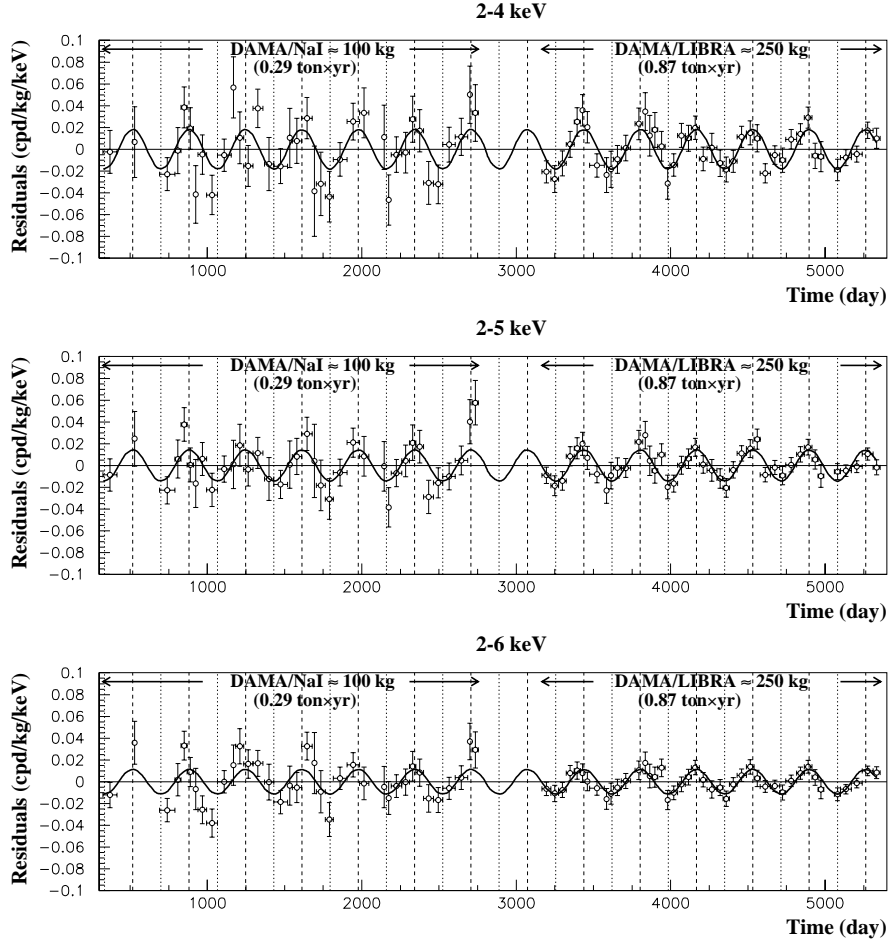
$$\Omega_{\text{PDM}} h^2 \sim \frac{G^{3/2} T_0^3 h^2}{H_0^2 \langle \sigma_{\text{ann}} v_{\text{rel}} \rangle} = \frac{3 \times 10^{-27} \text{ cm}^3 \text{ s}^{-1}}{\langle \sigma_{\text{ann}} v_{\text{rel}} \rangle}. \quad (116)$$

Here  $v_{\text{rel}}$  is the relative velocity of the two incoming dark matter particles and the brackets  $\langle \cdot \rangle$  denote a thermal average at the freeze-out temperature,  $T_f \simeq m_{\text{PDM}}/20$ , when the dark matter particles go out of equilibrium with radiation. The value of  $\langle \sigma_{\text{ann}} v_{\text{rel}} \rangle$  needed for  $\Omega_{\text{PDM}} \approx 1$  is remarkably close to what one would expect for a WIMP with a mass  $m_{\text{PDM}} = 100$  GeV,  $\langle \sigma_{\text{ann}} v_{\text{rel}} \rangle \sim \alpha^2/8\pi m_{\text{PDM}} \sim 3 \times 10^{-27} \text{ cm}^3 \text{ s}^{-1}$ . We still do not know whether this is just a coincidence or an important hint on the nature of dark matter.

There are a few theoretical candidates for WIMPs, like the neutralino, coming from supersymmetric extensions of the standard model of particle physics,<sup>6</sup> but at present there is no empirical evidence that

<sup>6</sup>For a review of Supersymmetry (SUSY), see Kazakov's contribution to these Proceedings.



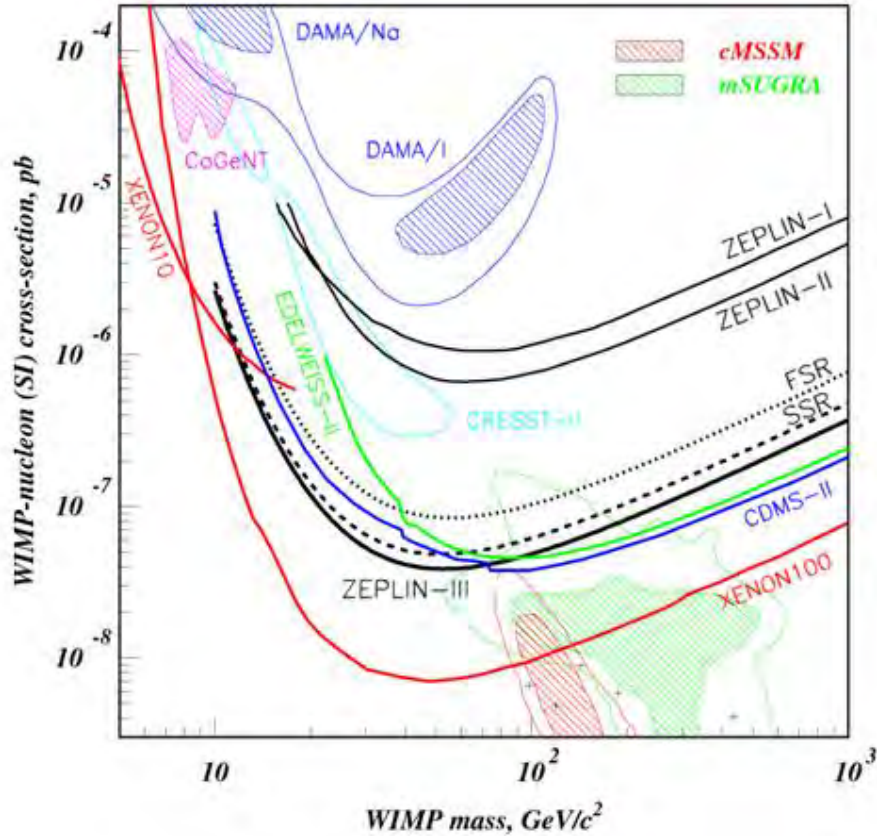


**Fig. 19:** The annual-modulation signal accumulated over 12 years is consistent with a neutralino of mass of  $m_\chi = 59^{+17}_{-14}$  GeV and a proton cross section of  $\xi\sigma_p = 7.0^{+0.4}_{-1.2} \times 10^{-6}$  pb, according to DAMA. From Ref. [55].

such extensions are indeed realized in nature. In fact, the non-observation of supersymmetric particles at current accelerators places stringent limits on the neutralino mass and interaction cross section [56]. If WIMPs constitute the dominant component of the halo of our galaxy, it is expected that some may cross the Earth at a reasonable rate to be detected. The direct experimental search for them rely on elastic WIMP collisions with the nuclei of a suitable target. Dark matter WIMPs move at a typical galactic “virial” velocity of around 200 – 300 km/s, depending on the model. If their mass is in the range 10 – 100 GeV, the recoil energy of the nuclei in the elastic collision would be of order 10 keV. Therefore, one should be able to identify such energy depositions in a macroscopic sample of the target. There are at present three different methods: First, one could search for scintillation light in NaI crystals or in liquid xenon; second, search for an ionization signal in a semiconductor, typically a very pure germanium crystal; and third, use a cryogenic detector at 10 mK and search for a measurable temperature increase of the sample. The main problem with such a type of experiment is the low expected signal rate, with a typical number below 1 event/kg/day. To reduce natural radioactive contamination one must use extremely pure substances, and to reduce the background caused by cosmic rays requires that these experiments be located deeply underground.

The best limits on WIMP scattering cross sections come from some germanium experiments, like the Cryogenic Dark Matter Search (CDMS) collaboration at Stanford and the Soudan mine [57], as well as from the NaI scintillation detectors of the UK dark matter collaboration (UKDMC) in the Boulby salt

mine in England [58], and the DAMA experiment in the Gran Sasso laboratory in Italy [55]. Current experiments already touch the parameter space expected from supersymmetric particles, see Fig. 20, and therefore there is a chance that they actually discover the nature of the missing dark matter. The problem, of course, is to attribute a tentative signal unambiguously to galactic WIMPs rather than to some unidentified radioactive background.



**Fig. 20:** Exclusion range for the spin-independent WIMP scattering cross section per nucleon from the NaI experiments and the Ge detectors. The blue lines come from the CDMS experiment, which exclude the DAMA region at more than 3 sigma. Also shown in yellow and red is the range of expected counting rates for neutralinos in the MSSM. From Ref. [57].

One specific signature is the annual modulation which arises as the Earth moves around the Sun.<sup>7</sup> Therefore, the net speed of the Earth relative to the galactic dark matter halo varies, causing a modulation of the expected counting rate. The DAMA/NaI experiment has actually reported such a modulation signal, from the combined analysis of their 12-year data, see Fig. 19 and Ref. [55], which provides a confidence level of 99.6% for a neutralino mass of  $m_\chi = 52^{+10}_{-8}$  GeV and a proton cross section of  $\xi\sigma_p = 7.2^{+0.4}_{-0.9} \times 10^{-6}$  pb, where  $\xi = \rho_\chi/0.3 \text{ GeV cm}^{-3}$  is the local neutralino energy density in units of the galactic halo density. There has been no confirmation yet of this result from other dark matter search groups. In fact, the CDMS collaboration claims an exclusion of the DAMA region at the 3 sigma level, see Fig. 20. Hopefully in the near future we will have much better sensitivity at low masses from the Cryogenic Rare Event Search with Superconducting Thermometers (CRESST) experiment at Gran Sasso. The CRESST experiment [59] uses sapphire crystals as targets and a new

<sup>7</sup>The time scale of the Sun's orbit around the center of the galaxy is too large to be relevant in the analysis.

method to simultaneously measure the phonons and the scintillating light from particle interactions inside the crystal, which allows excellent background discrimination. Very recently there has been also the proposal of a completely new method based on a Superheated Droplet Detector (SDD), which claims to have already a similar sensitivity as the more standard methods described above, see Ref. [60].

There exist other *indirect* methods to search for galactic WIMPs [61]. Such particles could self-annihilate at a certain rate in the galactic halo, producing a potentially detectable background of high energy photons or antiprotons. The absence of such a background in both gamma ray satellites and the Alpha Matter Spectrometer [62] imposes bounds on their density in the halo. Alternatively, WIMPs traversing the solar system may interact with the matter that makes up the Earth or the Sun so that a small fraction of them will lose energy and be trapped in their cores, building up over the age of the universe. Their annihilation in the core would thus produce high energy neutrinos from the center of the Earth or from the Sun which are detectable by neutrino telescopes. In fact, SuperKamiokande already covers a large part of SUSY parameter space. In other words, neutrino telescopes are already competitive with direct search experiments. In particular, the AMANDA experiment at the South Pole [63], which has approximately  $10^3$  Cherenkov detectors several km deep in very clear ice, over a volume  $\sim 1 \text{ km}^3$ , is competitive with the best direct searches proposed. The advantages of AMANDA are also directional, since the arrays of Cherenkov detectors will allow one to reconstruct the neutrino trajectory and thus its source, whether it comes from the Earth or the Sun. AMANDA recently reported the detection of TeV neutrinos [63].

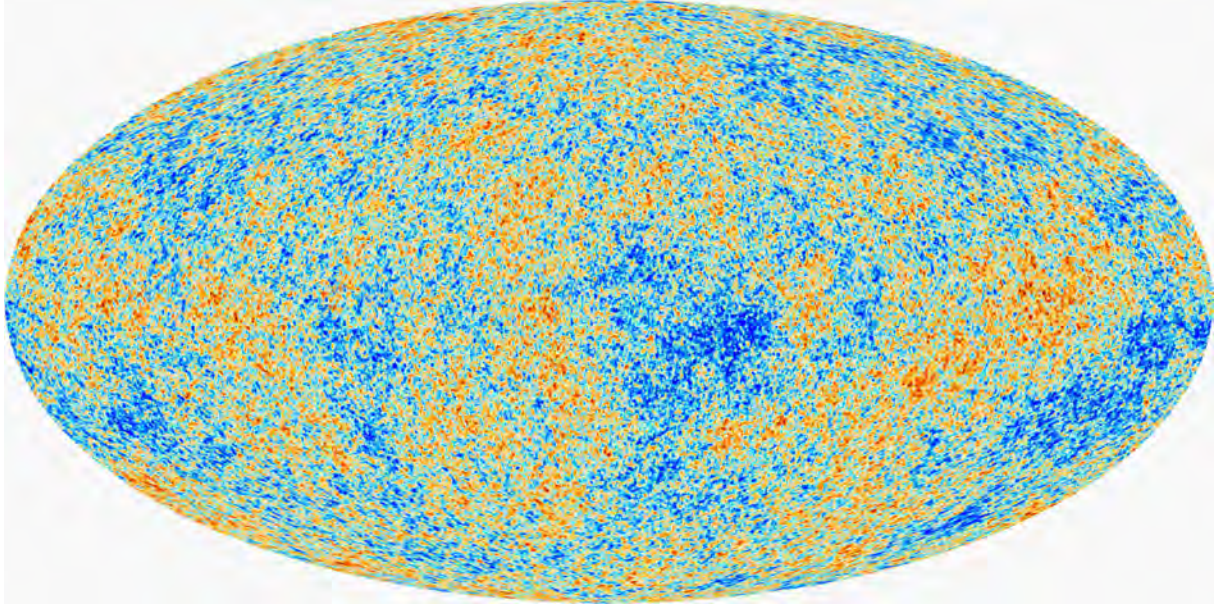
### 3.3 The age of the universe $t_0$

The universe must be older than the oldest objects it contains. Those are believed to be the stars in the oldest clusters in the Milky Way, globular clusters. The most reliable ages come from the application of theoretical models of stellar evolution to observations of old stars in globular clusters. For about 30 years, the ages of globular clusters have remained reasonable stable, at about 15 Gyr [64]. However, recently these ages have been revised downward [65].

During the 1980s and 1990s, the globular cluster age estimates have improved as both new observations have been made with CCDs, and since refinements to stellar evolution models, including opacities, consideration of mixing, and different chemical abundances have been incorporated [66]. From the theory side, uncertainties in globular cluster ages come from uncertainties in convection models, opacities, and nuclear reaction rates. From the observational side, uncertainties arise due to corrections for dust and chemical composition. However, the dominant source of systematic errors in the globular cluster age is the uncertainty in the cluster distances. Fortunately, the Hipparcos satellite recently provided geometric parallax measurements for many nearby old stars with low metallicity, typical of globular clusters, thus allowing for a new calibration of the ages of stars in globular clusters, leading to a downward revision to 10 – 13 Gyr [66]. Moreover, there were very few stars in the Hipparcos catalog with both small parallax errors and low metal abundance. Hence, an increase in the sample size could be critical in reducing the statistical uncertainties for the calibration of the globular cluster ages. There are already proposed two new parallax satellites, NASA's Space Interferometry Mission (SIM) and ESA's mission, called GAIA, that will give 2 or 3 orders of magnitude more accurate parallaxes than Hipparcos, down to fainter magnitude limits, for several orders of magnitude more stars. Until larger samples are available, however, distance errors are likely to be the largest source of systematic uncertainty to the globular cluster age [30].

The supernovae groups can also determine the age of the universe from their high redshift observations. The high confidence regions in the  $(\Omega_M, \Omega_\Lambda)$  plane are almost parallel to the contours of constant age. For any value of the Hubble constant less than  $H_0 = 70 \text{ km/s/Mpc}$ , the implied age of the universe is greater than 13 Gyr, allowing enough time for the oldest stars in globular clusters to evolve [66]. Integrating over  $\Omega_M$  and  $\Omega_\Lambda$ , the best fit value of the age in Hubble-time units is  $H_0 t_0 = 0.93 \pm 0.06$  or equivalently  $t_0 = 14.1 \pm 1.0 (0.65 h^{-1}) \text{ Gyr}$ , see Ref. [7]. Furthermore, a combination of 8 independent

recent measurements: CMB anisotropies, type Ia SNe, cluster mass-to-light ratios, cluster abundance evolution, cluster baryon fraction, deuterium-to-hydrogen ratios in quasar spectra, double-lobed radio sources and the Hubble constant, can be used to determine the present age of the universe [67]. The best fit value for the age of the universe is, according to this analysis,  $t_0 = 13.4 \pm 1.6$  Gyr, about a billion years younger than other recent estimates [67].

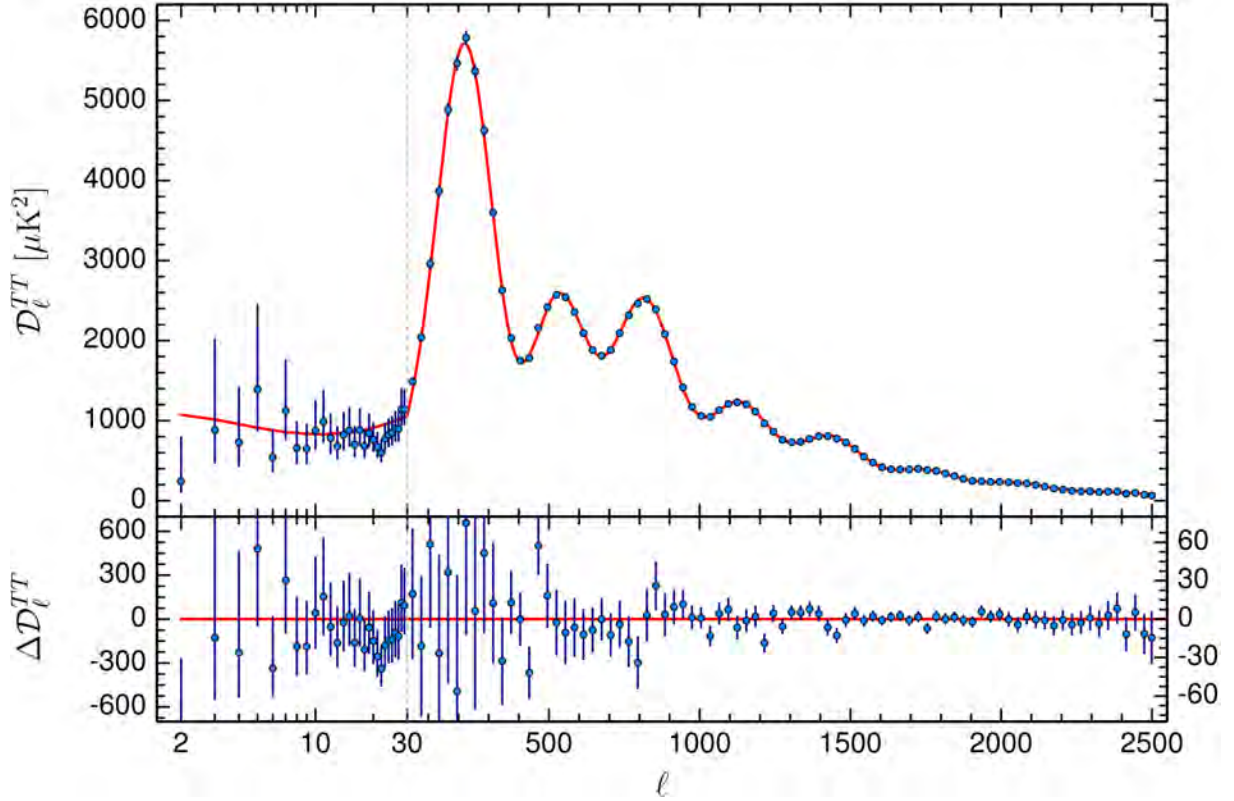


**Fig. 21:** The anisotropies of the microwave background measured by the Planck satellite with 4 arcminute resolution. It shows the intrinsic CMB anisotropies at the level of a few parts in  $10^5$ . The galactic foreground has been properly subtracted. The amount of information contained in this map is enough to determine most of the cosmological parameters to few percent accuracy. From Ref. [93].

### 3.4 Cosmic Microwave Background Anisotropies

The cosmic microwave background has become in the last five years the Holy Grail of Cosmology, since precise observations of the temperature and polarization anisotropies allow in principle to determine the parameters of the Standard Model of Cosmology with very high accuracy. Recently, the WMAP satellite has provided with a very detailed map of the microwave anisotropies in the sky, see Fig. 21, and indeed has fulfilled our expectations, see Table 2.

The physics of the CMB anisotropies is relatively simple [68]. The universe just before recombination is a very tightly coupled fluid, due to the large electromagnetic Thomson cross section  $\sigma_T = 8\pi\alpha^2/3m_e^2 \simeq 0.7$  barn. Photons scatter off charged particles (protons and electrons), and carry energy, so they feel the gravitational potential associated with the perturbations imprinted in the metric during inflation. An overdensity of baryons (protons and neutrons) does not collapse under the effect of gravity until it enters the causal Hubble radius. The perturbation continues to grow until radiation pressure opposes gravity and sets up acoustic oscillations in the plasma, very similar to sound waves. Since overdensities of the same size will enter the Hubble radius at the same time, they will oscillate in phase. Moreover, since photons scatter off these baryons, the acoustic oscillations occur also in the photon field and induces a pattern of peaks in the temperature anisotropies in the sky, at different angular scales, see Fig. 22. There are three different effects that determine the temperature anisotropies we observe in the CMB. First, *gravity*: photons fall in and escape off gravitational potential wells, characterized by  $\Phi$  in the comoving gauge, and as a consequence their frequency is gravitationally blue- or red-shifted,  $\delta\nu/\nu = \Phi$ . If the gravitational potential is not constant, the photons will escape from a larger or smaller potential



**Fig. 22:** The Angular Power Spectrum of CMB temperature anisotropies,  $\mathcal{D}_l = l(l+1)C_l/2\pi$ , and their residuals w.r.t. a Standard Cosmological Model with just six parameters. The low ( $l < 30$ ) multipoles are not used in the fit. The agreement is impressive, except for a small anomalous deviation around  $l \simeq 20$ . From Ref. [93].

well than they fell in, so their frequency is also blue- or red-shifted, a phenomenon known as the Rees-Sciama effect. Second, *pressure*: photons scatter off baryons which fall into gravitational potential wells and the two competing forces create acoustic waves of compression and rarefaction. Finally, *velocity*: baryons accelerate as they fall into potential wells. They have minimum velocity at maximum compression and rarefaction. That is, their velocity wave is exactly  $90^\circ$  off-phase with the acoustic waves. These waves induce a Doppler effect on the frequency of the photons. The temperature anisotropy induced by these three effects is therefore given by [68]

$$\frac{\delta T}{T}(\mathbf{r}) = \Phi(\mathbf{r}, t_{\text{dec}}) + 2 \int_{t_{\text{dec}}}^{t_0} \dot{\Phi}(\mathbf{r}, t) dt + \frac{1}{3} \frac{\delta \rho}{\rho} - \frac{\mathbf{r} \cdot \mathbf{v}}{c}. \quad (117)$$

Metric perturbations of different wavelengths enter the horizon at different times. The largest wavelengths, of size comparable to our present horizon, are entering now. There are perturbations with wavelengths comparable to the size of the horizon at the time of last scattering, of projected size about  $1^\circ$  in the sky today, which entered precisely at decoupling. And there are perturbations with wavelengths much smaller than the size of the horizon at last scattering, that entered much earlier than decoupling, all the way to the time of radiation-matter equality, which have gone through several acoustic oscillations before last scattering. All these perturbations of different wavelengths leave their imprint in the CMB anisotropies.

The baryons at the time of decoupling do not feel the gravitational attraction of perturbations with wavelength greater than the size of the horizon at last scattering, because of causality. Perturbations with exactly that wavelength are undergoing their first contraction, or acoustic compression, at decoupling. Those perturbations induce a large peak in the temperature anisotropies power spectrum, see Fig. 22.

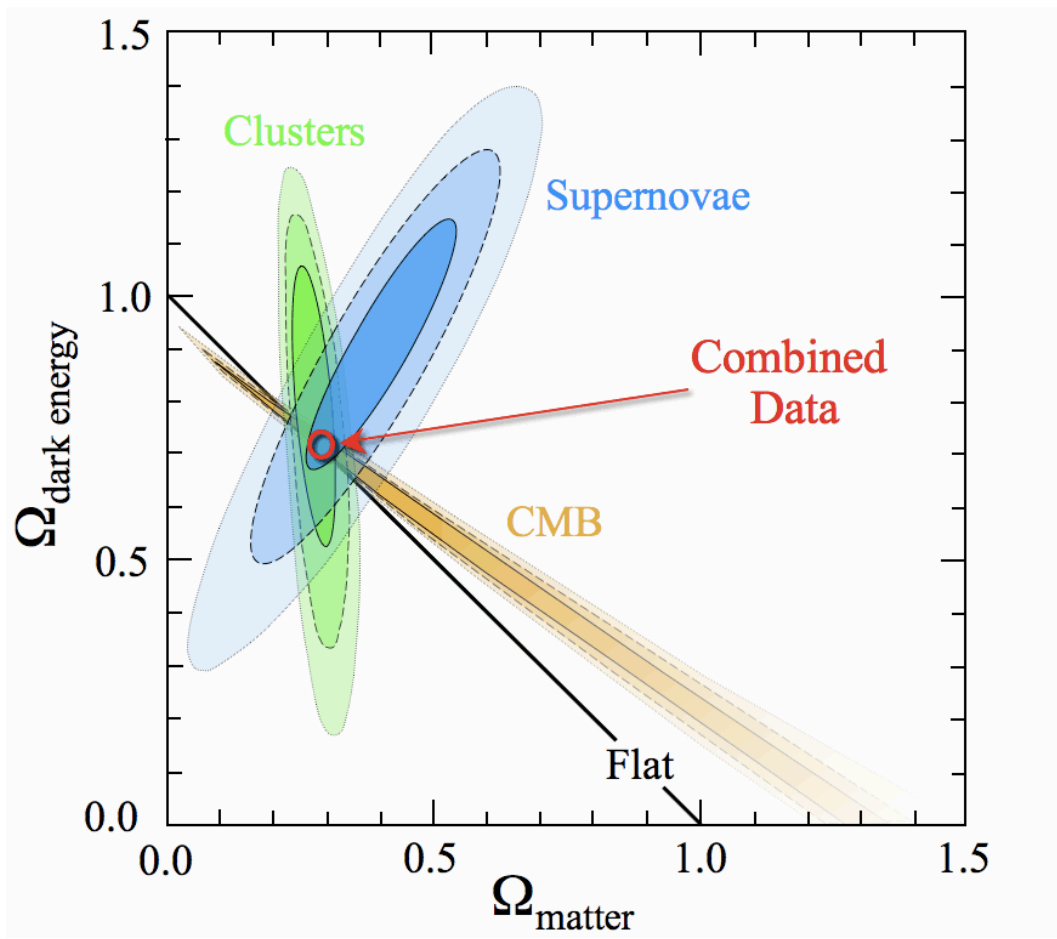
**Table 2: The parameters of the Standard Cosmological Model.** The standard model of cosmology has about 20 different parameters, needed to describe the background space-time, the matter content and the spectrum of metric perturbations. We include here the present range of the most relevant parameters (with  $1\sigma$  errors), as determined by both WMAP and Planck satellites. The rate of expansion is written in units of  $H = 100 h$  km/s/Mpc.

physical quantity	symbol	WMAP	Planck
total density	$\Omega_0$	$1.02 \pm 0.02$	$0.9992 \pm 0.0040$
baryonic matter	$\Omega_B h^2$	$0.0228 \pm 0.0020$	$0.02230 \pm 0.00014$
cosmological constant	$\Omega_\Lambda$	$0.73 \pm 0.04$	$0.6911 \pm 0.0062$
cold dark matter	$\Omega_M$	$0.23 \pm 0.04$	$0.3089 \pm 0.0062$
hot dark matter	$\Omega_\nu h^2$	$< 0.00276$ (95% c.l.)	$< 0.0021$ (95% c.l.)
number relativistic species	$N_{\text{eff}}$	$3.84 \pm 0.40$	$3.04 \pm 0.33$
sum of neutrino masses	$\sum m_\nu$ (eV)	$< 0.23$ (95% c.l.)	$< 0.194$ (95% c.l.)
CMB temperature	$T_0$ (K)	$2.725 \pm 0.002$	$2.7235 \pm 0.0012$
baryon to photon ratio	$10^{10} \eta$	$6.1 \pm 0.3$	$6.09 \pm 0.04$
baryon to matter ratio	$\Omega_B / \Omega_M$	$0.17 \pm 0.01$	$0.157 \pm 0.002$
spatial curvature	$\Omega_K$	$< 0.02$ (95% c.l.)	$0.0008 \pm 0.0040$
rate of expansion	$h$	$0.71 \pm 0.03$	$0.6774 \pm 0.0046$
age of the universe	$t_0$ (Gyr)	$13.7 \pm 0.6$	$13.799 \pm 0.021$
age at decoupling	$t_{\text{dec}}$ (kyr)	$379 \pm 8$	$380.000 \pm 0.120$
age at reionization	$t_r$ (Myr)	$180 \pm 100$	$212 \pm 43$
spectral amplitude	$10^{10} A_s^2$	$17.35 \pm 3.72$	$21.30 \pm 0.53$
spectral tilt	$n_s$	$0.98 \pm 0.03$	$0.9667 \pm 0.0040$
spectral tilt variation	$dn_s/d \ln k$	$-0.031 \pm 0.017$	$-0.002 \pm 0.013$
tensor-scalar ratio	$r$	$< 0.71$ (95% c.l.)	$< 0.113$ (95% c.l.)
reionization optical depth	$\tau$	$0.17 \pm 0.04$	$0.06 \pm 0.012$
redshift of equality	$z_{\text{eq}}$	$3233 \pm 200$	$3371 \pm 23$
redshift of decoupling	$z_{\text{dec}}$	$1089 \pm 1$	$1089.90 \pm 0.23$
width of decoupling	$\Delta z_{\text{dec}}$	$195 \pm 2$	$144.81 \pm 0.24$
redshift of reionization	$z_r$	$20 \pm 10$	$8.8 \pm 1.2$

Perturbations with wavelengths smaller than these will have gone, after they entered the Hubble scale, through a series of acoustic compressions and rarefactions, which can be seen as secondary peaks in the power spectrum. Since the surface of last scattering is not a sharp discontinuity, but a region of  $\Delta z \sim 100$ , there will be scales for which photons, traveling from one energy concentration to another, will erase the perturbation on that scale, similarly to what neutrinos or HDM do for structure on small scales. That is the reason why we don't see all the acoustic oscillations with the same amplitude, but in fact they decay exponentially towards smaller angular scales, an effect known as Silk damping, due to photon diffusion [68, 69].

From the observations of the CMB anisotropies it is possible to determine most of the parameters of the Standard Cosmological Model with few percent accuracy, see Table 2. However, there are many degeneracies between parameters and it is difficult to disentangle one from another. For instance, as mentioned above, the first peak in the photon distribution corresponds to overdensities that have undergone half an oscillation, that is, a compression, and appear at a scale associated with the size of the horizon at last scattering, about  $1^\circ$  projected in the sky today. Since photons scatter off baryons, they will also feel the acoustic wave and create a peak in the correlation function. The height of the peak is proportional to the amount of baryons: the larger the baryon content of the universe, the higher the

peak. The position of the peak in the power spectrum depends on the geometrical size of the particle horizon at last scattering. Since photons travel along geodesics, the projected size of the causal horizon at decoupling depends on whether the universe is flat, open or closed. In a flat universe the geodesics are straight lines and, by looking at the angular scale of the first acoustic peak, we would be measuring the actual size of the horizon at last scattering. In an open universe, the geodesics are inward-curved trajectories, and therefore the projected size on the sky appears smaller. In this case, the first acoustic peak should occur at higher multipoles or smaller angular scales. On the other hand, for a closed universe, the first peak occurs at smaller multipoles or larger angular scales. The dependence of the position of the first acoustic peak on the spatial curvature can be approximately given by  $l_{\text{peak}} \simeq 220 \Omega_0^{-1/2}$ , where  $\Omega_0 = \Omega_M + \Omega_\Lambda = 1 - \Omega_K$ . Present observations by WMAP and other experiments give  $\Omega_0 = 1.00 \pm 0.02$  at one standard deviation [21].



**Fig. 23:** The  $(\Omega_M, \Omega_\Lambda)$  plane with the present data set of cosmological observations – the acceleration of the universe, the large scale structure and the CMB anisotropies – on the fundamental parameters which define our Standard Model of Cosmology.

The other acoustic peaks occur at harmonics of this, corresponding to smaller angular scales. Since the amplitude and position of the primary and secondary peaks are directly determined by the sound speed (and, hence, the equation of state) and by the geometry and expansion of the universe, they can be used as a powerful test of the density of baryons and dark matter, and other cosmological parameters. With the joined data from WMAP, VSA, CBI and ACBAR, we have rather good evidence of the existence of the second and third acoustic peaks, which confirms one of the most important predictions of inflation – the non-causal origin of the primordial spectrum of perturbations –, and rules out cosmological defects

as the dominant source of structure in the universe [70]. Moreover, since the observations of CMB anisotropies now cover almost three orders of magnitude in the size of perturbations, we can determine the much better accuracy the value of the spectral tilt,  $n = 0.98 \pm 0.03$ , which is compatible with the approximate scale invariant spectrum needed for structure formation, and is a prediction of the simplest models of inflation. Soon after the release of data from WMAP, there was some expectation at the claim of a scale-dependent tilt. Nowadays, with better resolution in the linear matter power spectrum from SDSS [71], we can not conclude that the spectral tilt has any observable dependence on scale.

The microwave background has become also a testing ground for theories of particle physics. In particular, it already gives stringent constraints on the mass of the neutrino, when analysed together with large scale structure observations. Assuming a flat  $\Lambda$ CDM model, the 2-sigma upper bounds on the sum of the masses of light neutrinos is  $\sum m_\nu < 1.0$  eV for degenerate neutrinos (i.e. without a large hierarchy between them) if we don't impose any priors, and it comes down to  $\sum m_\nu < 0.6$  eV if one imposes the bounds coming from the HST measurements of the rate of expansion and the supernova data on the present acceleration of the universe [72]. The final bound on the neutrino density can be expressed as  $\Omega_\nu h^2 = \sum m_\nu / 93.2 \text{ eV} \leq 0.01$ . In the future, both with Planck and with the Atacama Cosmology Telescope (ACT) we will be able to put constraints on the neutrino masses down to the 0.1 eV level.

Moreover, the present data is good enough that we can start to put constraints on the models of inflation that give rise to structure. In particular, multifield models of inflation predict a mixture of adiabatic and isocurvature perturbations,<sup>8</sup> and their signatures in the cosmic microwave background anisotropies and the matter power spectrum of large scale structure are specific and perfectly distinguishable. Nowadays, thanks to precise CMB, LSS and SNIa data, one can put rather stringent limits on the relative fraction and correlation of the isocurvature modes to the dominant adiabatic perturbations [73].

We can summarize this Section by showing the region in parameter space where we stand nowadays, thanks to the recent cosmological observations. We have plotted that region in Fig. 23. One could also superimpose the contour lines corresponding to equal  $t_0 H_0$  lines, as a cross check. It is extraordinary that only in the last few months we have been able to reduce the concordance region to where it stands today, where all the different observations seem to converge. There are still many uncertainties, mainly systematic; however, those are quickly decreasing and becoming predominantly statistical. In the near future, with precise observations of the anisotropies in the microwave background temperature and polarization anisotropies, thanks to Planck satellite, we will be able to reduce those uncertainties to the level of one percent. This is the reason why cosmologists are so excited and why it is claimed that we live in the Golden Age of Cosmology.

## 4 The Inflationary Paradigm

The hot Big Bang theory is nowadays a very robust edifice, with many independent observational checks: the expansion of the universe; the abundance of light elements; the cosmic microwave background; a predicted age of the universe compatible with the age of the oldest objects in it, and the formation of structure via gravitational collapse of initially small inhomogeneities. Today, these observations are confirmed to within a few percent accuracy, and have helped establish the hot Big Bang as the preferred model of the universe. All the physics involved in the above observations is routinely tested in the laboratory (atomic and nuclear physics experiments) or in the solar system (general relativity).

However, this theory leaves a range of crucial questions unanswered, most of which are initial conditions' problems. There is the reasonable assumption that these cosmological problems will be solved or explained by *new physical principles* at high energies, in the early universe. This assumption leads to the natural conclusion that accurate observations of the present state of the universe may shed light onto processes and physical laws at energies above those reachable by particle accelerators, present

---

<sup>8</sup>This mixture is generic, unless all the fields thermalize simultaneously at reheating, just after inflation, in which case the entropy perturbations that would give rise to the isocurvature modes disappear.



or future. We will see that this is a very optimistic approach indeed, and that there are many unresolved issues related to those problems. However, there might be in the near future reasons to be optimistic.

#### 4.1 Shortcomings of Big Bang Cosmology

The Big Bang theory could not explain the origin of matter and structure in the universe; that is, the origin of the matter–antimatter asymmetry, without which the universe today would be filled by a uniform radiation continuously expanding and cooling, with no traces of matter, and thus without the possibility to form gravitationally bound systems like galaxies, stars and planets that could sustain life. Moreover, the standard Big Bang theory assumes, but cannot explain, the origin of the extraordinary smoothness and flatness of the universe on the very large scales seen by the microwave background probes and the largest galaxy catalogs. It cannot explain the origin of the primordial density perturbations that gave rise to cosmic structures like galaxies, clusters and superclusters, via gravitational collapse; the quantity and nature of the dark matter that we believe holds the universe together; nor the origin of the Big Bang itself.

A summary [10] of the problems that the Big Bang theory cannot explain is:

- The global structure of the universe.
  - Why is the universe so close to spatial flatness?
  - Why is matter so homogeneously distributed on large scales?
- The origin of structure in the universe.
  - How did the primordial spectrum of density perturbations originate?
- The origin of matter and radiation.
  - Where does all the energy in the universe come from?
  - What is the nature of the dark matter in the universe?
  - How did the matter-antimatter asymmetry arise?
- The initial singularity.
  - Did the universe have a beginning?
  - What is the global structure of the universe beyond our observable patch?

Let me discuss one by one the different issues:

##### 4.1.1 The Flatness Problem

The Big Bang theory assumes but cannot explain the extraordinary spatial flatness of our local patch of the universe. In the general FRW metric (2) the parameter  $K$  that characterizes spatial curvature is a free parameter. There is nothing in the theory that determines this parameter a priori. However, it is directly related, via the Friedmann equation (8), to the dynamics, and thus the matter content, of the universe,

$$K = \frac{8\pi G}{3}\rho a^2 - H^2 a^2 = \frac{8\pi G}{3}\rho a^2 \left( \frac{\Omega - 1}{\Omega} \right). \quad (118)$$

We can therefore define a new variable,

$$x \equiv \frac{\Omega - 1}{\Omega} = \frac{\text{const.}}{\rho a^2}, \quad (119)$$

whose time evolution is given by

$$x' = \frac{dx}{dN} = (1 + 3\omega) x, \quad (120)$$

where  $N = \ln(a/a_i)$  characterizes the *number of e-folds* of universe expansion ( $dN = Hdt$ ) and where we have used Eq. (7) for the time evolution of the total energy,  $\rho a^3$ , which only depends on the barotropic ratio  $\omega$ . It is clear from Eq. (120) that the phase-space diagram  $(x, x')$  presents an unstable

critical (saddle) point at  $x = 0$  for  $\omega > -1/3$ , i.e. for the radiation ( $\omega = 1/3$ ) and matter ( $\omega = 0$ ) eras. A small perturbation from  $x = 0$  will drive the system towards  $x = \pm\infty$ . Since we know the universe went through both the radiation era (because of primordial nucleosynthesis) and the matter era (because of structure formation), tiny deviations from  $\Omega = 1$  would have grown since then, such that today

$$x_0 = \frac{\Omega_0 - 1}{\Omega_0} = x_{\text{in}} \left( \frac{T_{\text{in}}}{T_{\text{eq}}} \right)^2 (1 + z_{\text{eq}}). \quad (121)$$

In order that today's value be in the range  $0.1 < \Omega_0 < 1.2$ , or  $x_0 \approx \mathcal{O}(1)$ , it is required that at, say, primordial nucleosynthesis ( $T_{\text{NS}} \simeq 10^6 T_{\text{eq}}$ ) its value be

$$\Omega(t_{\text{NS}}) = 1 \pm 10^{-15}, \quad (122)$$

which represents a tremendous finetuning. Perhaps the universe indeed started with such a peculiar initial condition, but it is epistemologically more satisfying if we give a fundamental dynamical reason for the universe to have started so close to spatial flatness. These arguments were first used by Robert Dicke in the 1960s, much before inflation. He argued that the most natural initial condition for the spatial curvature should have been the Planck scale curvature,  ${}^{(3)}R = 6K/l_{\text{P}}^2$ , where the Planck length is  $l_{\text{P}} = (\hbar G/c^3)^{1/2} = 1.62 \times 10^{-33}$  cm, that is, 60 orders of magnitude smaller than the present size of the universe,  $a_0 = 1.38 \times 10^{28}$  cm. A universe with this immense curvature would have collapsed within a Planck time,  $t_{\text{P}} = (\hbar G/c^5)^{1/2} = 5.39 \times 10^{-44}$  s, again 60 orders of magnitude smaller than the present age of the universe,  $t_0 = 4.1 \times 10^{17}$  s. Therefore, the flatness problem is also related to the Age Problem, why is it that the universe is so old and flat when, under ordinary circumstances (based on the fundamental scale of gravity) it should have lasted only a Planck time and reached a size of order the Planck length? As we will see, inflation gives a dynamical reason to such a peculiar initial condition.

#### 4.1.2 The Homogeneity Problem

An expanding universe has *particle horizons*, that is, spatial regions beyond which causal communication cannot occur. The horizon distance can be defined as the maximum distance that light could have travelled since the origin of the universe [16],

$$d_{\text{H}}(t) \equiv a(t) \int_0^t \frac{dt'}{a(t')} \sim H^{-1}(t), \quad (123)$$

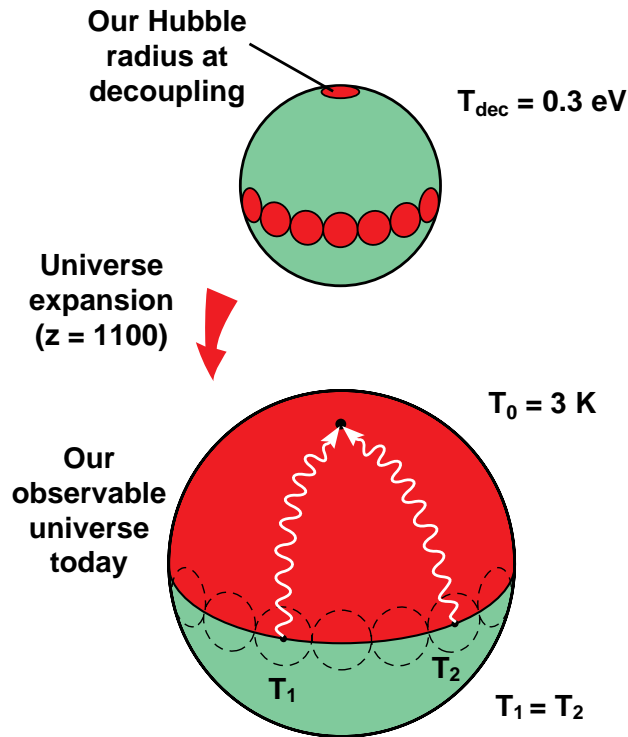
which is proportional to the Hubble scale.<sup>9</sup> For instance, at the beginning of nucleosynthesis the horizon distance is a few light-seconds, but grows *linearly* with time and by the end of nucleosynthesis it is a few light-minutes, i.e. a factor 100 larger, while the scale factor has increased *only* a factor of 10. The fact that the causal horizon increases faster,  $d_{\text{H}} \sim t$ , than the scale factor,  $a \sim t^{1/2}$ , implies that at any given time the universe contains regions within itself that, according to the Big Bang theory, were *never* in causal contact before. For instance, the number of causally disconnected regions at a given redshift  $z$  present in our causal volume today,  $d_{\text{H}}(t_0) \equiv a_0$ , is

$$N_{\text{CD}}(z) \sim \left( \frac{a(t)}{d_{\text{H}}(t)} \right)^3 \simeq (1+z)^{3/2}, \quad (124)$$

which, for the time of decoupling, is of order  $N_{\text{CD}}(z_{\text{dec}}) \sim 10^5 \gg 1$ .

This phenomenon is particularly acute in the case of the observed microwave background. Information cannot travel faster than the speed of light, so the causal region at the time of photon decoupling could not be larger than  $d_{\text{H}}(t_{\text{dec}}) \sim 3 \times 10^5$  light years across, or about  $1^\circ$  projected in the sky today. So why should regions that are separated by more than  $1^\circ$  in the sky today have exactly the same temperature, to within 10 ppm, when the photons that come from those two distant regions could not have been in causal contact when they were emitted? This constitutes the so-called horizon problem, see Fig. 24, and was first discussed by Robert Dicke in the 1970s as a profound inconsistency of the Big Bang theory.

<sup>9</sup>For the radiation era, the horizon distance is equal to the Hubble scale. For the matter era it is twice the Hubble scale.

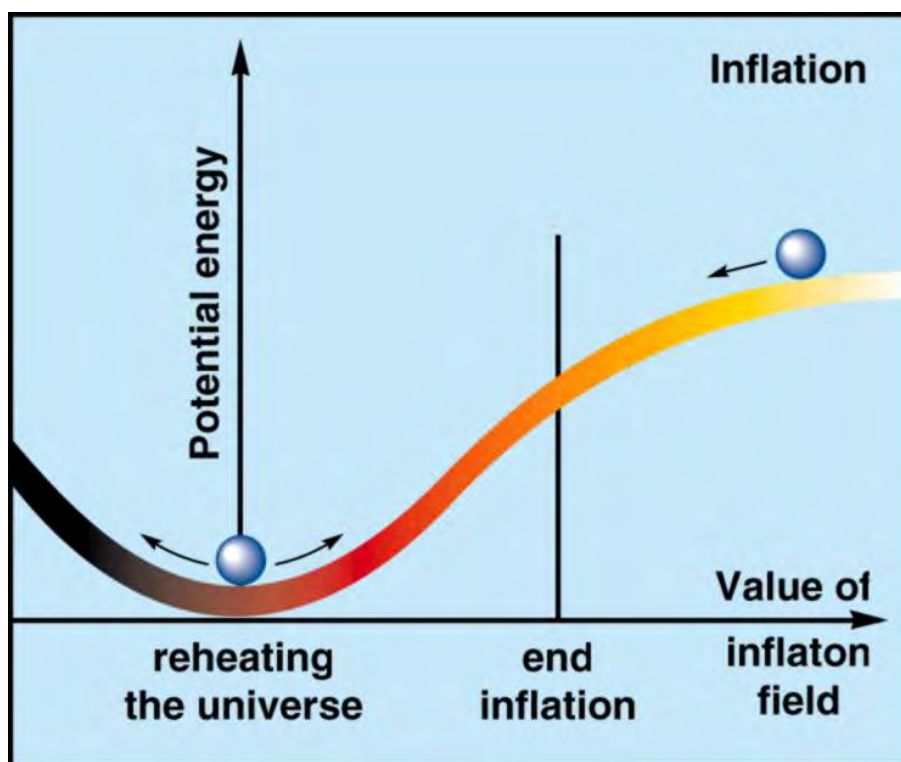


**Fig. 24:** Perhaps the most acute problem of the Big Bang theory is explaining the extraordinary homogeneity and isotropy of the microwave background, see Fig. 9. At the time of decoupling, the volume that gave rise to our present universe contained many causally disconnected regions (top figure). Today we observe a blackbody spectrum of photons coming from those regions and they appear to have the same temperature,  $T_1 = T_2$ , to one part in  $10^5$ . Why is the universe so homogeneous? This constitutes the so-called horizon problem, which is spectacularly solved by inflation. From Ref. [74].

### 4.2 Cosmological Inflation

In the 1980s, a new paradigm, deeply rooted in fundamental physics, was put forward by Alan H. Guth [75], Andrei D. Linde [76] and others [77–79], to address these fundamental questions. According to the inflationary paradigm, the early universe went through a period of exponential expansion, driven by the approximately constant energy density of a scalar field called the inflaton. In modern physics, elementary particles are represented by quantum fields, which resemble the familiar electric, magnetic and gravitational fields. A field is simply a function of space and time whose quantum oscillations are interpreted as particles. In our case, the inflaton field has, associated with it, a large potential energy density, which drives the exponential expansion during inflation, see Fig. 25. We know from general relativity that the density of matter determines the expansion of the universe, but a constant energy density acts in a very peculiar way: as a repulsive force that makes any two points in space separate at exponentially large speeds. (This does not violate the laws of causality because there is no information carried along in the expansion, it is simply the stretching of space-time.)

This superluminal expansion is capable of explaining the large scale homogeneity of our observable universe and, in particular, why the microwave background looks so isotropic: regions separated today by more than  $1^\circ$  in the sky were, in fact, in causal contact before inflation, but were stretched to

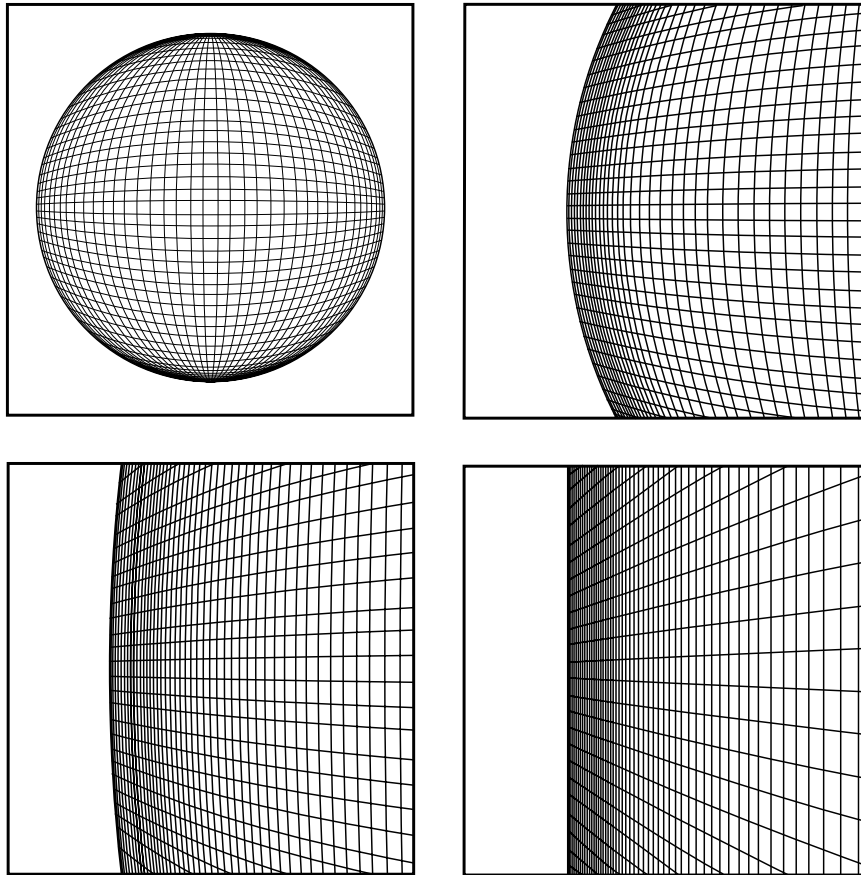


**Fig. 25:** The inflaton field can be represented as a ball rolling down a hill. During inflation, the energy density is approximately constant, driving the tremendous expansion of the universe. When the ball starts to oscillate around the bottom of the hill, inflation ends and the inflaton energy decays into particles. In certain cases, the coherent oscillations of the inflaton could generate a resonant production of particles which soon thermalize, reheating the universe. From Ref. [74].

cosmological distances by the expansion. Any inhomogeneities present before the tremendous expansion would be washed out. This explains why photons from supposedly causally disconnected regions have actually the same spectral distribution with the same temperature, see Fig. 24.

Moreover, in the usual Big Bang scenario a flat universe, one in which the gravitational attraction of matter is exactly balanced by the cosmic expansion, is unstable under perturbations: a small deviation from flatness is amplified and soon produces either an empty universe or a collapsed one. As we discussed above, for the universe to be nearly flat today, it must have been extremely flat at nucleosynthesis, deviations not exceeding more than one part in  $10^{15}$ . This extreme fine tuning of initial conditions was also solved by the inflationary paradigm, see Fig. 26. Thus inflation is an extremely elegant hypothesis that explains how a region much, much greater than our own observable universe could have become smooth and flat without recourse to *ad hoc* initial conditions. Furthermore, inflation dilutes away any “unwanted” relic species that could have remained from early universe phase transitions, like monopoles, cosmic strings, etc., which are predicted in grand unified theories and whose energy density could be so large that the universe would have become unstable, and collapsed, long ago. These relics are diluted by the superluminal expansion, which leaves at most one of these particles per causal horizon, making them harmless to the subsequent evolution of the universe.

The only thing we know about this peculiar scalar field, the *inflaton*, is that it has a mass and a self-interaction potential  $V(\phi)$  but we ignore everything else, even the scale at which its dynamics



**Fig. 26:** The exponential expansion during inflation made the radius of curvature of the universe so large that our observable patch of the universe today appears essentially flat, analogous (in three dimensions) to how the surface of a balloon appears flatter and flatter as we inflate it to enormous sizes. This is a crucial prediction of cosmological inflation that will be tested to extraordinary accuracy in the next few years. From Ref. [74, 78].

determines the superluminal expansion. In particular, we still do not know the nature of the inflaton field itself, is it some new *fundamental* scalar field in the electroweak symmetry breaking sector, or is it just some *effective* description of a more fundamental high energy interaction? Hopefully, in the near future, experiments in particle physics might give us a clue to its nature. Inflation had its original inspiration in the Higgs field, the scalar field supposed to be responsible for the masses of elementary particles (quarks and leptons) and the breaking of the electroweak symmetry. Such a field has not been found yet, and its discovery at the future particle colliders would help understand one of the truly fundamental problems in physics, the origin of masses. If the experiments discover something completely new and unexpected, it would automatically affect the idea of inflation at a fundamental level.

#### 4.2.1 Homogeneous scalar field dynamics

In this subsection I will describe the theoretical basis for the phenomenon of inflation. Consider a scalar field  $\phi$ , a singlet under any given interaction, with an effective potential  $V(\phi)$ . The Lagrangian for such a field in a curved background is

$$\mathcal{L}_{\text{inf}} = \frac{1}{2} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - V(\phi), \quad (125)$$

whose evolution equation in a Friedmann-Robertson-Walker metric (2) and for a *homogeneous* field  $\phi(t)$  is given by

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0, \quad (126)$$

where  $H$  is the rate of expansion, together with the Einstein equations,

$$H^2 = \frac{\kappa^2}{3} \left( \frac{1}{2} \dot{\phi}^2 + V(\phi) \right), \quad (127)$$

$$\dot{H} = -\frac{\kappa^2}{2} \dot{\phi}^2, \quad (128)$$

where  $\kappa^2 \equiv 8\pi G$ . The dynamics of inflation can be described as a perfect fluid (5) with a time dependent pressure and energy density given by

$$\rho = \frac{1}{2} \dot{\phi}^2 + V(\phi), \quad (129)$$

$$p = \frac{1}{2} \dot{\phi}^2 - V(\phi). \quad (130)$$

The field evolution equation (126) can then be written as the energy conservation equation,

$$\dot{\rho} + 3H(\rho + p) = 0. \quad (131)$$

If the potential energy density of the scalar field dominates the kinetic energy,  $V(\phi) \gg \dot{\phi}^2$ , then we see that

$$p \simeq -\rho \quad \Rightarrow \quad \rho \simeq \text{const.} \quad \Rightarrow \quad H(\phi) \simeq \text{const.}, \quad (132)$$

which leads to the solution

$$a(t) \sim \exp(Ht) \quad \Rightarrow \quad \frac{\ddot{a}}{a} > 0 \quad \text{accelerated expansion.} \quad (133)$$

Using the definition of the number of  $e$ -folds,  $N = \ln(a/a_i)$ , we see that the scale factor grows exponentially,  $a(N) = a_i \exp(N)$ . This solution of the Einstein equations solves immediately the flatness problem. Recall that the problem with the radiation and matter eras is that  $\Omega = 1$  ( $x = 0$ ) is an unstable critical point in phase-space. However, during inflation, with  $p \simeq -\rho \Rightarrow \omega \simeq -1$ , we have that  $1 + 3\omega \geq 0$  and therefore  $x = 0$  is a stable *attractor* of the equations of motion, see Eq. (120). As a consequence, what seemed an *ad hoc* initial condition, becomes a natural *prediction* of inflation. Suppose that during inflation the scale factor increased  $N$   $e$ -folds, then

$$x_0 = x_{\text{in}} e^{-2N} \left( \frac{T_{\text{rh}}}{T_{\text{eq}}} \right)^2 (1 + z_{\text{eq}}) \simeq e^{-2N} 10^{56} \leq 1 \quad \Rightarrow \quad N \geq 65, \quad (134)$$

where we have assumed that inflation ended at the scale  $V_{\text{end}}$ , and the transfer of the inflaton energy density to thermal radiation at reheating occurred almost instantaneously<sup>10</sup> at the temperature  $T_{\text{rh}} \sim V_{\text{end}}^{1/4} \sim 10^{15}$  GeV. Note that we can now have initial conditions with a large uncertainty,  $x_{\text{in}} \simeq 1$ , and still have today  $x_0 \simeq 1$ , thanks to the inflationary attractor towards  $\Omega = 1$ . This can be understood very easily by realizing that the three curvature evolves during inflation as

$${}^{(3)}R = \frac{6K}{a^2} = {}^{(3)}R_{\text{in}} e^{-2N} \quad \longrightarrow \quad 0, \quad \text{for } N \gg 1. \quad (135)$$

Therefore, if cosmological inflation lasted over 65  $e$ -folds, as most models predict, then today the universe (or at least our local patch) should be exactly flat, see Fig. 26, a prediction that can be tested with

<sup>10</sup>There could be a small delay in thermalization, due to the intrinsic inefficiency of reheating, but this does not change significantly the required number of  $e$ -folds.

great accuracy in the near future and for which already seems to be some evidence from observations of the microwave background [91].

Furthermore, inflation also solves the homogeneity problem in a spectacular way. First of all, due to the superluminal expansion, any inhomogeneity existing prior to inflation will be washed out,

$$\delta_k \sim \left( \frac{k}{aH} \right)^2 \Phi_k \propto e^{-2N} \longrightarrow 0, \quad \text{for } N \gg 1. \quad (136)$$

Moreover, since the scale factor grows exponentially, while the horizon distance remains essentially constant,  $d_H(t) \simeq H^{-1} = \text{const.}$ , any scale within the horizon during inflation will be stretched by the superluminal expansion to enormous distances, in such a way that at photon decoupling all the causally disconnected regions that encompass our present horizon actually come from a single region during inflation, about 65  $e$ -folds before the end. This is the reason why two points separated more than  $1^\circ$  in the sky have the same backbody temperature, as observed by the COBE satellite: they were actually in causal contact during inflation. There is at present no other proposal known that could solve the homogeneity problem without invoquing an acausal mechanism like inflation.

Finally, any relic particle species (relativistic or not) existing prior to inflation will be diluted by the expansion,

$$\rho_M \propto a^{-3} \sim e^{-3N} \longrightarrow 0, \quad \text{for } N \gg 1, \quad (137)$$

$$\rho_R \propto a^{-4} \sim e^{-4N} \longrightarrow 0, \quad \text{for } N \gg 1. \quad (138)$$

Note that the vacuum energy density  $\rho_v$  remains constant under the expansion, and therefore, very soon it is the only energy density remaining to drive the expansion of the universe.

#### 4.2.2 The slow-roll approximation

In order to simplify the evolution equations during inflation, we will consider the slow-roll approximation (SRA). Suppose that, during inflation, the scalar field evolves very slowly down its effective potential, then we can define the slow-roll parameters [80],

$$\epsilon \equiv -\frac{\dot{H}}{H^2} = \frac{\kappa^2 \dot{\phi}^2}{2 H^2} \ll 1, \quad (139)$$

$$\delta \equiv -\frac{\ddot{\phi}}{H\dot{\phi}} \ll 1, \quad (140)$$

$$\xi \equiv \frac{\dddot{\phi}}{H^2\dot{\phi}} - \delta^2 \ll 1. \quad (141)$$

It is easy to see that the condition

$$\epsilon < 1 \iff \frac{\ddot{a}}{a} > 0 \quad (142)$$

characterizes inflation: it is all you need for superluminal expansion, i.e. for the horizon distance to grow more slowly than the scale factor, in order to solve the homogeneity problem, as well as for the spatial curvature to decay faster than usual, in order to solve the flatness problem.

The number of  $e$ -folds during inflation can be written with the help of Eq. (139) as

$$N = \ln \frac{a_{\text{end}}}{a_i} = \int_{t_i}^{t_e} H dt = \int_{\phi_i}^{\phi_e} \frac{\kappa d\phi}{\sqrt{2\epsilon(\phi)}}, \quad (143)$$

which is an exact expression in terms of  $\epsilon(\phi)$ .

In the limit given by Eqs. (139), the evolution equations (126) and (127) become

$$H^2 \left(1 - \frac{\epsilon}{3}\right) \simeq H^2 = \frac{\kappa^2}{3} V(\phi), \quad (144)$$

$$3H\dot{\phi} \left(1 - \frac{\delta}{3}\right) \simeq 3H\dot{\phi} = -V'(\phi). \quad (145)$$

Note that this corresponds to a reduction of the dimensionality of phase-space from two to one dimensions,  $H(\phi, \dot{\phi}) \rightarrow H(\phi)$ . In fact, it is possible to prove a theorem, for single-field inflation, which states that the slow-roll approximation is an attractor of the equations of motion, and thus we can always evaluate the inflationary trajectory in phase-space within the SRA, therefore reducing the number of initial conditions to just one, the initial value of the scalar field. If  $H(\phi)$  only depends on  $\phi$ , then  $H'(\phi) = -\kappa^2 \dot{\phi}/2$  and we can rewrite the slow-roll parameters (139) as

$$\epsilon = \frac{2}{\kappa^2} \left( \frac{H'(\phi)}{H(\phi)} \right)^2 \simeq \frac{1}{2\kappa^2} \left( \frac{V'(\phi)}{V(\phi)} \right)^2 \equiv \epsilon_V \ll 1, \quad (146)$$

$$\delta = \frac{2}{\kappa^2} \frac{H''(\phi)}{H(\phi)} \simeq \frac{1}{\kappa^2} \frac{V''(\phi)}{V(\phi)} - \frac{1}{2\kappa^2} \left( \frac{V'(\phi)}{V(\phi)} \right)^2 \equiv \eta_V - \epsilon_V \ll 1, \quad (147)$$

$$\begin{aligned} \xi &= \frac{4}{\kappa^4} \frac{H'(\phi)H'''(\phi)}{H^2(\phi)} \simeq \frac{1}{\kappa^4} \frac{V'(\phi)V'''(\phi)}{V^2(\phi)} - \frac{3}{2\kappa^4} \frac{V''(\phi)}{V(\phi)} \left( \frac{V'(\phi)}{V(\phi)} \right)^2 \\ &+ \frac{3}{4\kappa^4} \left( \frac{V'(\phi)}{V(\phi)} \right)^4 \equiv \xi_V - 3\eta_V\epsilon_V + 3\epsilon_V^2 \ll 1. \end{aligned} \quad (148)$$

These expressions define the new slow-roll parameters  $\epsilon_V$ ,  $\eta_V$  and  $\xi_V$ . The number of  $e$ -folds can also be rewritten in this approximation as

$$N \simeq \int_{\phi_i}^{\phi_e} \frac{\kappa d\phi}{\sqrt{2\epsilon_V(\phi)}} = \kappa^2 \int_{\phi_i}^{\phi_e} \frac{V(\phi) d\phi}{V'(\phi)}, \quad (149)$$

a very useful expression for evaluating  $N$  for a given effective scalar potential  $V(\phi)$ .

### 4.3 The origin of density perturbations

If cosmological inflation made the universe so extremely flat and homogeneous, where did the galaxies and clusters of galaxies come from? One of the most astonishing predictions of inflation, one that was not even expected, is that quantum fluctuations of the inflaton field are stretched by the exponential expansion and generate large-scale perturbations in the metric. Inflaton fluctuations are small wave packets of energy that, according to general relativity, modify the space-time fabric, creating a whole spectrum of curvature perturbations. The use of the word spectrum here is closely related to the case of light waves propagating in a medium: a spectrum characterizes the amplitude of each given wavelength. In the case of inflation, the inflaton fluctuations induce waves in the space-time metric that can be decomposed into different wavelengths, all with approximately the same amplitude, that is, corresponding to a scale-invariant spectrum. These patterns of perturbations in the metric are like fingerprints that unequivocally characterize a period of inflation. When matter fell in the troughs of these waves, it created density perturbations that collapsed gravitationally to form galaxies, clusters and superclusters of galaxies, with a spectrum that is also scale invariant. Such a type of spectrum was proposed in the early 1970s (before inflation) by Harrison and Zel'dovich [27], to explain the distribution of galaxies and clusters of galaxies on very large scales in our observable universe. Perhaps the most interesting aspect of structure formation is the possibility that the detailed knowledge of what seeded galaxies and clusters of galaxies will allow us to test the idea of inflation.



### 4.3.1 Reparametrization invariant perturbation theory

Until now we have considered only the unperturbed FRW metric described by a scale factor  $a(t)$  and a homogeneous scalar field  $\phi(t)$ ,

$$ds^2 = a^2(\eta)[-d\eta^2 + \gamma_{ij}dx^i dx^j], \quad (150)$$

$$\phi = \phi(\eta), \quad (151)$$

where  $\eta = \int dt/a(t)$  is the conformal time, under which the background equations of motion can be written as

$$\mathcal{H}^2 = \frac{\kappa^2}{3} \left( \frac{1}{2} \phi'^2 + a^2 V(\phi) \right), \quad (152)$$

$$\mathcal{H}' - \mathcal{H}^2 = -\frac{\kappa^2}{2} \phi'^2, \quad (153)$$

$$\phi'' + 2\mathcal{H}\phi' + a^2 V'(\phi) = 0, \quad (154)$$

where  $\mathcal{H} = aH$  and  $\phi' = a\dot{\phi}$ .

During inflation, the quantum fluctuations of the scalar field will induce metric perturbations which will backreact on the scalar field. Let us consider, in linear perturbation theory, the most general line element with both scalar and tensor metric perturbations [81],<sup>11</sup> together with the scalar field perturbations

$$ds^2 = a^2(\eta) \left[ -(1 + 2A)d\eta^2 + 2B_{|i}dx^i d\eta + \left\{ (1 + 2\mathcal{R})\gamma_{ij} + 2E_{|ij} + 2h_{ij} \right\} dx^i dx^j \right], \quad (155)$$

$$\phi = \phi(\eta) + \delta\phi(\eta, x^i). \quad (156)$$

The indices  $\{i, j\}$  label the three-dimensional spatial coordinates with metric  $\gamma_{ij}$ , and the  $|i$  denotes covariant derivative with respect to that metric. The gauge invariant tensor perturbation  $h_{ij}$  corresponds to a transverse traceless gravitational wave,  $\nabla^i h_{ij} = h_i^i = 0$ . The four scalar perturbations ( $A, B, \mathcal{R}, E$ ) are *gauge dependent* functions of  $(\eta, x^i)$ . Under a general coordinate (gauge) transformation [81, 82]

$$\tilde{\eta} = \eta + \xi^0(\eta, x^i), \quad (157)$$

$$\tilde{x}^i = x^i + \gamma^{ij}\xi_{|j}(\eta, x^i), \quad (158)$$

with arbitrary functions  $(\xi^0, \xi)$ , the scalar and tensor perturbations transform, to linear order, as

$$\tilde{A} = A - \xi^{0'} - \mathcal{H}\xi^0, \quad \tilde{B} = B + \xi^0 - \xi', \quad (159)$$

$$\tilde{\mathcal{R}} = \mathcal{R} - \mathcal{H}\xi^0, \quad \tilde{E} = E - \xi, \quad (160)$$

$$\tilde{h}_{ij} = h_{ij}, \quad (161)$$

where a prime denotes derivative with respect to conformal time. It is possible to construct, however, two gauge-invariant gravitational potentials [81, 82],

$$\Phi = A + (B - E')' + \mathcal{H}(B - E'), \quad (162)$$

$$\Psi = \mathcal{R} + \mathcal{H}(B - E'), \quad (163)$$

which are related through the perturbed Einstein equations,

$$\Phi = \Psi, \quad (164)$$

$$\frac{k^2 - 3K}{a^2} \Psi = \frac{\kappa^2}{2} \delta\rho, \quad (165)$$

<sup>11</sup>Note that inflation cannot generate, to linear order, a vector perturbation.

where  $\delta\rho$  is the gauge-invariant density perturbation, and the latter expression is nothing but the Poisson equation for the gravitational potential, written in relativistic form.

During inflation, the energy density is given in terms of a scalar field, and thus the gauge-invariant equations for the perturbations on comoving hypersurfaces (constant energy density hypersurfaces) are

$$\Phi'' + 3\mathcal{H}\Phi' + (\mathcal{H}' + 2\mathcal{H}^2)\Phi = \frac{\kappa^2}{2}[\phi'\delta\phi' - a^2V'(\phi)\delta\phi], \quad (166)$$

$$-\nabla^2\Phi + 3\mathcal{H}\Phi' + (\mathcal{H}' + 2\mathcal{H}^2)\Phi = -\frac{\kappa^2}{2}[\phi'\delta\phi' + a^2V'(\phi)\delta\phi], \quad (167)$$

$$\Phi' + \mathcal{H}\Phi = \frac{\kappa^2}{2}\phi'\delta\phi, \quad (168)$$

$$\delta\phi'' + 2\mathcal{H}\delta\phi' - \nabla^2\delta\phi = 4\phi'\Phi' - 2a^2V'(\phi)\Phi - a^2V''(\phi)\delta\phi. \quad (169)$$

This system of equations seem too difficult to solve at first sight. However, there is a gauge invariant combination of variables that allows one to find exact solutions. Let us define [82]

$$u \equiv a\delta\phi + z\Phi, \quad (170)$$

$$z \equiv a\frac{\phi'}{\mathcal{H}}. \quad (171)$$

Under this redefinition, the above equations simplify enormously to just three independent equations,

$$u'' - \nabla^2u - \frac{z''}{z}u = 0, \quad (172)$$

$$\nabla^2\Phi = \frac{\kappa^2}{2}\frac{\mathcal{H}}{a^2}(zu' - z'u), \quad (173)$$

$$\left(\frac{a^2\Phi}{\mathcal{H}}\right)' = \frac{\kappa^2}{2}zu. \quad (174)$$

From Equation (172) we can find a solution  $u(z)$ , which substituted into (174) can be integrated to give  $\Phi(z)$ , and together with  $u(z)$  allow us to obtain  $\delta\phi(z)$ .

### 4.3.2 Quantum Field Theory in curved space-time

Until now we have treated the perturbations as classical, but we should in fact consider the perturbations  $\Phi$  and  $\delta\phi$  as quantum fields. Note that the perturbed action for the scalar mode  $u$  can be written as

$$\delta S = \frac{1}{2} \int d^3x d\eta \left[ (u')^2 - (\nabla u)^2 + \frac{z''}{z}u^2 \right]. \quad (175)$$

In order to quantize the field  $u$  in the curved background defined by the metric (150), we can write the operator

$$\hat{u}(\eta, \mathbf{x}) = \int \frac{d^3\mathbf{k}}{(2\pi)^{3/2}} \left[ u_k(\eta) \hat{a}_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} + u_k^*(\eta) \hat{a}_{\mathbf{k}}^\dagger e^{-i\mathbf{k}\cdot\mathbf{x}} \right], \quad (176)$$

where the creation and annihilation operators satisfy the commutation relation of bosonic fields, and the scalar field's Fock space is defined through the vacuum condition,

$$[\hat{a}_{\mathbf{k}}, \hat{a}_{\mathbf{k}'}^\dagger] = \delta^3(\mathbf{k} - \mathbf{k}'), \quad (177)$$

$$\hat{a}_{\mathbf{k}}|0\rangle = 0. \quad (178)$$

Note that we are not assuming that the inflaton is a fundamental scalar field, but that it can be written as a quantum field with its commutation relations (as much as a pion can be described as a quantum field).

The equations of motion for each mode  $u_k(\eta)$  are decoupled in linear perturbation theory,

$$u_k'' + \left(k^2 - \frac{z''}{z}\right)u_k = 0. \quad (179)$$

The ratio  $z''/z$  acts like a time-dependent potential for this Schrödinger like equation. In order to find exact solutions to the mode equation, we will use the slow-roll parameters (139), see Ref. [80]

$$\epsilon = 1 - \frac{\mathcal{H}'}{\mathcal{H}^2} = \frac{\kappa^2 z^2}{2 a^2}, \quad (180)$$

$$\delta = 1 - \frac{\phi''}{\mathcal{H}\phi'} = 1 + \epsilon - \frac{z'}{\mathcal{H}z}, \quad (181)$$

$$\xi = - \left(2 - \epsilon - 3\delta + \delta^2 - \frac{\phi'''}{\mathcal{H}^2\phi'}\right). \quad (182)$$

In terms of these parameters, the conformal time and the effective potential for the  $u_k$  mode can be written as

$$\eta = \frac{-1}{\mathcal{H}} + \int \frac{\epsilon da}{a\mathcal{H}}, \quad (183)$$

$$\frac{z''}{z} = \mathcal{H}^2 \left[ (1 + \epsilon - \delta)(2 - \delta) + \mathcal{H}^{-1}(\epsilon' - \delta') \right]. \quad (184)$$

Note that the slow-roll parameters, (180) and (181), can be taken as *constant*,<sup>12</sup> to order  $\epsilon^2$ ,

$$\begin{aligned} \epsilon' &= 2\mathcal{H}(\epsilon^2 - \epsilon\delta) = \mathcal{O}(\epsilon^2), \\ \delta' &= \mathcal{H}(\epsilon\delta - \xi) = \mathcal{O}(\epsilon^2). \end{aligned} \quad (185)$$

In that case, for constant slow-roll parameters, we can write

$$\eta = \frac{-1}{\mathcal{H}} \frac{1}{1 - \epsilon}, \quad (186)$$

$$\frac{z''}{z} = \frac{1}{\eta^2} \left( \nu^2 - \frac{1}{4} \right), \quad \text{where} \quad \nu = \frac{1 + \epsilon - \delta}{1 - \epsilon} + \frac{1}{2}. \quad (187)$$

We are now going to search for approximate solutions of the mode equation (179), where the effective potential (184) is of order  $z''/z \simeq 2\mathcal{H}^2$  in the slow-roll approximation. In quasi-de Sitter there is a characteristic scale given by the (event) horizon size or Hubble scale during inflation,  $H^{-1}$ . There will be modes  $u_k$  with physical wavelengths much smaller than this scale,  $k/a \gg H$ , that are well within the de Sitter horizon and therefore do not feel the curvature of space-time. On the other hand, there will be modes with physical wavelengths much greater than the Hubble scale,  $k/a \ll H$ . In these two asymptotic regimes, the solutions can be written as

$$u_k = \frac{1}{\sqrt{2k}} e^{-ik\eta} \quad k \gg aH, \quad (188)$$

$$u_k = C_1 z \quad k \ll aH. \quad (189)$$

In the limit  $k \gg aH$  the modes behave like ordinary quantum modes in Minkowsky space-time, appropriately normalized, while in the opposite limit,  $u/z$  becomes constant on superhorizon scales. For

<sup>12</sup>For instance, there are models of inflation, like power-law inflation,  $a(t) \sim t^p$ , where  $\epsilon = \delta = 1/p < 1$ , that give constant slow-roll parameters.

approximately constant slow-roll parameters one can find exact solutions to (179), with the effective potential given by (187), that interpolate between the two asymptotic solutions,

$$u_k(\eta) = \frac{\sqrt{\pi}}{2} e^{i(\nu+\frac{1}{2})\frac{\pi}{2}} (-\eta)^{1/2} H_\nu^{(1)}(-k\eta), \quad (190)$$

where  $H_\nu^{(1)}(z)$  is the Hankel function of the first kind [83], and  $\nu$  is given by (187) in terms of the slow-roll parameters. In the limit  $k\eta \rightarrow 0$ , the solution becomes

$$|u_k| = \frac{2^{\nu-\frac{3}{2}}}{\sqrt{2k}} \frac{\Gamma(\nu)}{\Gamma(\frac{3}{2})} (-k\eta)^{\frac{1}{2}-\nu} \equiv \frac{C(\nu)}{\sqrt{2k}} \left(\frac{k}{aH}\right)^{\frac{1}{2}-\nu}, \quad (191)$$

$$C(\nu) = 2^{\nu-\frac{3}{2}} \frac{\Gamma(\nu)}{\Gamma(\frac{3}{2})} (1-\epsilon)^{\nu-\frac{1}{2}} \simeq 1 \quad \text{for } \epsilon, \delta \ll 1. \quad (192)$$

We can now compute  $\Phi$  and  $\delta\phi$  from the super-Hubble-scale mode solution (189), for  $k \ll aH$ . Substituting into Eq. (174), we find

$$\Phi = C_1 \left(1 - \frac{\mathcal{H}}{a^2} \int a^2 d\eta\right) + C_2 \frac{\mathcal{H}}{a^2}, \quad (193)$$

$$\delta\phi = \frac{C_1}{a^2} \int a^2 d\eta - \frac{C_2}{a^2}. \quad (194)$$

The term proportional to  $C_1$  corresponds to the growing solution, while that proportional to  $C_2$  corresponds to the decaying solution, which can soon be ignored. These quantities are gauge invariant but evolve with time outside the horizon, during inflation, and before entering again the horizon during the radiation or matter eras. We would like to write an expression for a gauge invariant quantity that is also *constant* for superhorizon modes. Fortunately, in the case of adiabatic perturbations, there is such a quantity:

$$\zeta \equiv \Phi + \frac{1}{\epsilon\mathcal{H}} (\Phi' + \mathcal{H}\Phi) = \frac{u}{z}, \quad (195)$$

which is constant, see Eq. (189), for  $k \ll aH$ . In fact, this quantity  $\zeta$  is identical, for superhorizon modes, to the gauge invariant curvature metric perturbation  $\mathcal{R}_c$  on comoving (constant energy density) hypersurfaces, see Ref. [81, 84],

$$\zeta = \mathcal{R}_c + \frac{1}{\epsilon\mathcal{H}^2} \nabla^2 \Phi. \quad (196)$$

Using Eq. (173) we can write the evolution equation for  $\zeta = \frac{u}{z}$  as  $\zeta' = \frac{1}{\epsilon\mathcal{H}} \nabla^2 \Phi$ , which confirms that  $\zeta$  is constant for (adiabatic<sup>13</sup>) superhorizon modes,  $k \ll aH$ . Therefore, we can evaluate the Newtonian potential  $\Phi_k$  when the perturbation reenters the horizon during radiation/matter eras in terms of the curvature perturbation  $\mathcal{R}_k$  when it left the Hubble scale during inflation,

$$\Phi_k = \left(1 - \frac{\mathcal{H}}{a^2} \int a^2 d\eta\right) \mathcal{R}_k = \frac{3+3\omega}{5+3\omega} \mathcal{R}_k = \begin{cases} \frac{2}{3} \mathcal{R}_k & \text{radiation era,} \\ \frac{3}{5} \mathcal{R}_k & \text{matter era.} \end{cases} \quad (197)$$

Let us now compute the tensor or gravitational wave metric perturbations generated during inflation. The perturbed action for the tensor mode can be written as

$$\delta S = \frac{1}{2} \int d^3x d\eta \frac{a^2}{2\kappa^2} \left[ (h'_{ij})^2 - (\nabla h_{ij})^2 \right], \quad (198)$$

<sup>13</sup>This conservation fails for entropy or isocurvature perturbations, see Ref. [84].

with the tensor field  $h_{ij}$  considered as a quantum field,

$$\hat{h}_{ij}(\eta, \mathbf{x}) = \int \frac{d^3\mathbf{k}}{(2\pi)^{3/2}} \sum_{\lambda=1,2} \left[ h_k(\eta) e_{ij}(\mathbf{k}, \lambda) \hat{a}_{\mathbf{k},\lambda} e^{i\mathbf{k}\cdot\mathbf{x}} + h.c. \right], \quad (199)$$

where  $e_{ij}(\mathbf{k}, \lambda)$  are the two polarization tensors, satisfying symmetric, transverse and traceless conditions

$$e_{ij} = e_{ji}, \quad k^i e_{ij} = 0, \quad e_{ii} = 0, \quad (200)$$

$$e_{ij}(-\mathbf{k}, \lambda) = e_{ij}^*(\mathbf{k}, \lambda), \quad \sum_{\lambda} e_{ij}^*(\mathbf{k}, \lambda) e^{ij}(\mathbf{k}, \lambda) = 4, \quad (201)$$

while the creation and annihilation operators satisfy the usual commutation relation of bosonic fields, Eq. (177). We can now redefine our gauge invariant tensor amplitude as

$$v_k(\eta) = \frac{a}{\sqrt{2k}} h_k(\eta), \quad (202)$$

which satisfies the following evolution equation, decoupled for each mode  $v_k(\eta)$  in linear perturbation theory,

$$v_k'' + \left( k^2 - \frac{a''}{a} \right) v_k = 0. \quad (203)$$

The ratio  $a''/a$  acts like a time-dependent potential for this Schrödinger like equation, analogous to the term  $z''/z$  for the scalar metric perturbation. For constant slow-roll parameters, the potential becomes

$$\frac{a''}{a} = 2\mathcal{H}^2 \left( 1 - \frac{\epsilon}{2} \right) = \frac{1}{\eta^2} \left( \mu^2 - \frac{1}{4} \right), \quad (204)$$

$$\mu = \frac{1}{1-\epsilon} + \frac{1}{2}. \quad (205)$$

We can solve equation (203) in the two asymptotic regimes,

$$v_k = \frac{1}{\sqrt{2k}} e^{-ik\eta} \quad k \gg aH, \quad (206)$$

$$v_k = C a \quad k \ll aH. \quad (207)$$

In the limit  $k \gg aH$  the modes behave like ordinary quantum modes in Minkowsky space-time, appropriately normalized, while in the opposite limit, the metric perturbation  $h_k$  becomes *constant* on superhorizon scales. For constant slow-roll parameters one can find exact solutions to (203), with effective potential given by (204), that interpolate between the two asymptotic solutions. These are identical to Eq. (190) except for the substitution  $\nu \rightarrow \mu$ . In the limit  $k\eta \rightarrow 0$ , the solution becomes

$$|v_k| = \frac{C(\mu)}{\sqrt{2k}} \left( \frac{k}{aH} \right)^{\frac{1}{2}-\mu}. \quad (208)$$

Since the mode  $h_k$  becomes constant on superhorizon scales, we can evaluate the tensor metric perturbation when it reentered during the radiation or matter era directly in terms of its value during inflation.

### 4.3.3 Power spectrum of scalar and tensor metric perturbations

Not only do we expect to measure the amplitude of the metric perturbations generated during inflation and responsible for the anisotropies in the CMB and density fluctuations in LSS, but we should also be able to measure its power spectrum, or two-point correlation function in Fourier space. Let us consider

first the scalar metric perturbations  $\mathcal{R}_k$ , which enter the horizon at  $a = k/H$ . Its correlator is given by [80]

$$\langle 0 | \mathcal{R}_k^* \mathcal{R}_{k'} | 0 \rangle = \frac{|u_k|^2}{z^2} \delta^3(\mathbf{k} - \mathbf{k}') \equiv \frac{\mathcal{P}_{\mathcal{R}}(k)}{4\pi k^3} (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}'), \quad (209)$$

$$\mathcal{P}_{\mathcal{R}}(k) = \frac{k^3}{2\pi^2} \frac{|u_k|^2}{z^2} = \frac{\kappa^2}{2\epsilon} \left( \frac{H}{2\pi} \right)^2 \left( \frac{k}{aH} \right)^{3-2\nu} \equiv A_S^2 \left( \frac{k}{aH} \right)^{n_s-1}, \quad (210)$$

where we have used  $\mathcal{R}_k = \zeta_k = \frac{u_k}{z}$  and Eq. (191). This last equation determines the power spectrum in terms of its amplitude at horizon-crossing,  $A_S$ , and a tilt,

$$n_s - 1 \equiv \frac{d \ln \mathcal{P}_{\mathcal{R}}(k)}{d \ln k} = 3 - 2\nu = 2 \left( \frac{\delta - 2\epsilon}{1 - \epsilon} \right) \simeq 2\eta_V - 6\epsilon_V, \quad (211)$$

see Eqs. (146), (147). Note from this equation that it is possible, in principle, to obtain from inflation a scalar tilt which is either positive ( $n > 1$ ) or negative ( $n < 1$ ). Furthermore, depending on the particular inflationary model [85], we can have significant departures from scale invariance.

Note that at horizon entry  $k\eta = -1$ , and thus we can alternatively evaluate the tilt as

$$n_s - 1 \equiv -\frac{d \ln \mathcal{P}_{\mathcal{R}}}{d \ln \eta} = -2\eta\mathcal{H} \left[ (1 - \epsilon) - (\epsilon - \delta) - 1 \right] = 2 \left( \frac{\delta - 2\epsilon}{1 - \epsilon} \right) \simeq 2\eta_V - 6\epsilon_V, \quad (212)$$

and the running of the tilt

$$\frac{dn_s}{d \ln k} = -\frac{dn_s}{d \ln \eta} = -\eta\mathcal{H} \left( 2\xi + 8\epsilon^2 - 10\epsilon\delta \right) \simeq 2\xi_V + 24\epsilon_V^2 - 16\eta_V\epsilon_V, \quad (213)$$

where we have used Eqs. (185).

Let us consider now the tensor (gravitational wave) metric perturbation, which enter the horizon at  $a = k/H$ ,

$$\sum_{\lambda} \langle 0 | h_{k,\lambda}^* h_{k',\lambda} | 0 \rangle = 4 \frac{2\kappa^2}{a^2} |v_k|^2 \delta^3(\mathbf{k} - \mathbf{k}') \equiv \frac{\mathcal{P}_g(k)}{4\pi k^3} (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}'), \quad (214)$$

$$\mathcal{P}_g(k) = 8\kappa^2 \left( \frac{H}{2\pi} \right)^2 \left( \frac{k}{aH} \right)^{3-2\mu} \equiv A_T^2 \left( \frac{k}{aH} \right)^{n_T}, \quad (215)$$

where we have used Eqs. (202) and (208). Therefore, the power spectrum can be approximated by a power-law expression, with amplitude  $A_T$  and tilt

$$n_T \equiv \frac{d \ln \mathcal{P}_g(k)}{d \ln k} = 3 - 2\mu = \frac{-2\epsilon}{1 - \epsilon} \simeq -2\epsilon_V < 0, \quad (216)$$

which is always negative. In the slow-roll approximation,  $\epsilon \ll 1$ , the tensor power spectrum is scale invariant.

Alternatively, we can evaluate the tensor tilt by

$$n_T \equiv -\frac{d \ln \mathcal{P}_g}{d \ln \eta} = -2\eta\mathcal{H} \left[ (1 - \epsilon) - 1 \right] = \frac{-2\epsilon}{1 - \epsilon} \simeq -2\epsilon_V, \quad (217)$$

and its running by

$$\frac{dn_T}{d \ln k} = -\frac{dn_T}{d \ln \eta} = -\eta\mathcal{H} \left( 4\epsilon^2 - 4\epsilon\delta \right) \simeq 8\epsilon_V^2 - 4\eta_V\epsilon_V, \quad (218)$$

where we have used Eqs. (185).

#### 4.4 The anisotropies of the microwave background

The metric fluctuations generated during inflation are not only responsible for the density perturbations that gave rise to galaxies via gravitational collapse, but one should also expect to see such ripples in the metric as temperature anisotropies in the cosmic microwave background, that is, minute deviations in the temperature of the blackbody spectrum when we look at different directions in the sky. Such anisotropies had been looked for ever since Penzias and Wilson's discovery of the CMB, but had eluded all detection, until COBE satellite discovered them in 1992, see Fig. 9. The reason why they took so long to be discovered was that they appear as perturbations in temperature of only one part in  $10^5$ . Soon after COBE, other groups quickly confirmed the detection of temperature anisotropies at around  $30 \mu\text{K}$ , at higher multipole numbers or smaller angular scales.

##### 4.4.1 The Sachs-Wolfe effect

The anisotropies corresponding to large angular scales are only generated via gravitational red-shift and density perturbations through the Einstein equations,  $\delta\rho/\rho = -2\Phi$  for adiabatic perturbations; we can ignore the Doppler contribution, since the perturbation is non-causal. In that case, the temperature anisotropy in the sky today is given by [86]

$$\frac{\delta T}{T}(\theta, \phi) = \frac{1}{3}\Phi(\eta_{\text{LS}}) Q(\eta_0, \theta, \phi) + 2 \int_{\eta_{\text{LS}}}^{\eta_0} dr \Phi'(\eta_0 - r) Q(r, \theta, \phi), \quad (219)$$

where  $\eta_0$  is the *coordinate distance* to the last scattering surface, i.e. the present conformal time, while  $\eta_{\text{LS}} \simeq 0$  determines that comoving hypersurface. The above expression is known as the Sachs-Wolfe effect [86], and contains two parts, the intrinsic and the Integrated Sachs-Wolfe (ISW) effect, due to integration along the line of sight of time variations in the gravitational potential.

In linear perturbation theory, the scalar metric perturbations can be separated into  $\Phi(\eta, \mathbf{x}) \equiv \Phi(\eta) Q(\mathbf{x})$ , where  $Q(\mathbf{x})$  are the scalar harmonics, eigenfunctions of the Laplacian in three dimensions,  $\nabla^2 Q_{klm}(r, \theta, \phi) = -k^2 Q_{klm}(r, \theta, \phi)$ . These functions have the general form [87]

$$Q_{klm}(r, \theta, \phi) = \Pi_{kl}(r) Y_{lm}(\theta, \phi), \quad (220)$$

where  $Y_{lm}(\theta, \phi)$  are the usual spherical harmonics [83].

In order to compute the temperature anisotropy associated with the Sachs-Wolfe effect, we have to know the evolution of the metric perturbation during the matter era,

$$\Phi'' + 3\mathcal{H}\Phi' + a^2\Lambda\Phi - 2K\Phi = 0. \quad (221)$$

In the case of a flat universe without cosmological constant, the Newtonian potential remains constant during the matter era and only the intrinsic SW effect contributes to  $\delta T/T$ . In case of a non-vanishing  $\Lambda$ , since its contribution is negligible in the past, most of the photon's trajectory towards us is unperturbed, and the only difference with respect to the  $\Lambda = 0$  case is an overall factor [90]. We will consider here the approximation  $\Phi = \text{constant}$  during the matter era and ignore that factor, see Ref. [88].

In a flat universe, the radial part of the eigenfunctions (220) can be written as [87]

$$\Pi_{kl}(r) = \sqrt{\frac{2}{\pi}} k j_l(kr), \quad (222)$$

where  $j_l(z)$  are the spherical Bessel functions [83]. The growing mode solution of the metric perturbation that left the Hubble scale during inflation contributes to the temperature anisotropies on large scales (219) as

$$\frac{\delta T}{T}(\theta, \phi) = \frac{1}{3}\Phi(\eta_{\text{LS}}) Q = \frac{1}{5}\mathcal{R} Q(\eta_0, \theta, \phi) \equiv \sum_{l=2}^{\infty} \sum_{m=-l}^l a_{lm} Y_{lm}(\theta, \phi), \quad (223)$$

where we have used the fact that at reentry (at the surface of last scattering) the gauge invariant Newtonian potential  $\Phi$  is related to the curvature perturbation  $\mathcal{R}$  at Hubble-crossing during inflation, see Eq. (197); and we have expanded  $\delta T/T$  in spherical harmonics.

We can now compute the two-point correlation function or angular power spectrum,  $C(\theta)$ , of the CMB anisotropies on large scales, defined as an expansion in multipole number,

$$C(\theta) = \left\langle \frac{\delta T^*}{T}(\mathbf{n}) \frac{\delta T}{T}(\mathbf{n}') \right\rangle_{\mathbf{n} \cdot \mathbf{n}' = \cos \theta} = \frac{1}{4\pi} \sum_{l=2}^{\infty} (2l+1) C_l P_l(\cos \theta), \quad (224)$$

where  $P_l(z)$  are the Legendre polynomials [83], and we have averaged over different universe realizations. Since the coefficients  $a_{lm}$  are isotropic (to first order), we can compute the  $C_l = \langle |a_{lm}|^2 \rangle$  as

$$C_l^{(S)} = \frac{4\pi}{25} \int_0^{\infty} \frac{dk}{k} \mathcal{P}_{\mathcal{R}}(k) j_l^2(k\eta_0), \quad (225)$$

where we have used Eqs. (223) and (209). In the case of scalar metric perturbation produced during inflation, the scalar power spectrum at reentry is given by  $\mathcal{P}_{\mathcal{R}}(k) = A_S^2 (k\eta_0)^{n-1}$ , in the power-law approximation, see Eq. (210). In that case, one can integrate (225) to give

$$C_l^{(S)} = \frac{2\pi}{25} A_S^2 \frac{\Gamma[\frac{3}{2}] \Gamma[1 - \frac{n-1}{2}] \Gamma[l + \frac{n-1}{2}]}{\Gamma[\frac{3}{2} - \frac{n-1}{2}] \Gamma[l + 2 - \frac{n-1}{2}]}, \quad (226)$$

$$\frac{l(l+1) C_l^{(S)}}{2\pi} = \frac{A_S^2}{25} = \text{constant}, \quad \text{for } n = 1. \quad (227)$$

This last expression corresponds to what is known as the Sachs-Wolfe plateau, and is the reason why the coefficients  $C_l$  are always plotted multiplied by  $l(l+1)$ , see Fig. 3.4.

Tensor metric perturbations also contribute with an approximately constant angular power spectrum,  $l(l+1)C_l$ . The Sachs-Wolfe effect for a gauge invariant tensor perturbation is given by [86]

$$\frac{\delta T}{T}(\theta, \phi) = \int_{\eta_{\text{LS}}}^{\eta_0} dr h'(\eta_0 - r) Q_{rr}(r, \theta, \phi), \quad (228)$$

where  $Q_{rr}$  is the  $rr$ -component of the tensor harmonic along the line of sight [87]. The tensor perturbation  $h$  during the matter era satisfies the following evolution equation

$$h_k'' + 3\mathcal{H} h_k' + (k^2 + 2K) h_k = 0, \quad (229)$$

which depends on the wavenumber  $k$ , contrary to what happens with the scalar modes, see Eq. (221). For a flat ( $K = 0$ ) universe, the solution to this equation is  $h_k(\eta) = h G_k(\eta)$ , where  $h$  is the constant tensor metric perturbation at horizon crossing and  $G_k(\eta) = 3 j_1(k\eta)/k\eta$ , normalized so that  $G_k(0) = 1$  at the surface of last scattering. The radial part of the tensor harmonic  $Q_{rr}$  in a flat universe can be written as [87]

$$Q_{kl}^{rr}(r) = \left[ \frac{(l-1)l(l+1)(l+2)}{\pi k^2} \right]^{1/2} \frac{j_l(kr)}{r^2}. \quad (230)$$

The tensor angular power spectrum can finally be expressed as

$$C_l^{(T)} = \frac{9\pi}{4} (l-1)l(l+1)(l+2) \int_0^{\infty} \frac{dk}{k} \mathcal{P}_g(k) I_{kl}^2, \quad (231)$$

$$I_{kl} = \int_0^{x_0} dx \frac{j_2(x_0 - x) j_l(x)}{(x_0 - x)x^2}, \quad (232)$$



where  $x \equiv k\eta$ , and  $\mathcal{P}_g(k)$  is the primordial tensor spectrum (215). For a scale invariant spectrum,  $n_T = 0$ , we can integrate (231) to give [89]

$$l(l+1)C_l^{(T)} = \frac{\pi}{36} \left(1 + \frac{48\pi^2}{385}\right) A_T^2 B_l, \quad (233)$$

with  $B_l = (1.1184, 0.8789, \dots, 1.00)$  for  $l = 2, 3, \dots, 30$ . Therefore,  $l(l+1)C_l^{(T)}$  also becomes constant for large  $l$ . Beyond  $l \sim 30$ , the Sachs-Wolfe expression is not a good approximation and the tensor angular power spectrum decays very quickly at large  $l$ , see Fig. 28.

#### 4.4.2 The consistency relation

In spite of the success of inflation in predicting a homogeneous and isotropic background on which to imprint a scale-invariant spectrum of inhomogeneities, it is difficult to test the idea of inflation. A CMB cosmologist before the 1980s would have argued that *ad hoc* initial conditions could have been at the origin of the homogeneity and flatness of the universe on large scales, while a LSS cosmologist would have agreed with Harrison and Zel'dovich that the most natural spectrum needed to explain the formation of structure was a scale-invariant spectrum. The surprise was that inflation incorporated an understanding of *both* the globally homogeneous and spatially flat background, and the approximately scale-invariant spectrum of perturbations in the same formalism. But that could have been just a coincidence.

What is *unique* to inflation is the fact that inflation determines not just one but *two* primordial spectra, corresponding to the scalar (density) and tensor (gravitational waves) metric perturbations, from a single continuous function, the inflaton potential  $V(\phi)$ . In the slow-roll approximation, one determines, from  $V(\phi)$ , two continuous functions,  $\mathcal{P}_{\mathcal{R}}(k)$  and  $\mathcal{P}_g(k)$ , that in the power-law approximation reduces to two amplitudes,  $A_S$  and  $A_T$ , and two tilts,  $n$  and  $n_T$ . It is clear that there must be a relation between the four parameters. Indeed, one can see from Eqs. (233) and (227) that the ratio of the tensor to scalar contribution to the angular power spectrum is proportional to the tensor tilt [80],

$$r \equiv \frac{A_T^2}{A_S^2} = 16\epsilon \simeq -8n_T. \quad (234)$$

This is a unique prediction of inflation, which could not have been postulated a priori by any cosmologist. If we finally observe a tensor spectrum of anisotropies in the CMB, or a stochastic gravitational wave background in laser interferometers like LIGO or LISA, with sufficient accuracy to determine their spectral tilt, one might have some chance to test the idea of inflation, via the consistency relation (234). For the moment, observations of the microwave background anisotropies suggest that the Sachs-Wolfe plateau exists, see Fig. 3.4, but it is still premature to determine the tensor contribution. Perhaps in the near future, from the analysis of polarization as well as temperature anisotropies, with the CMB satellites MAP and Planck, we might have a chance of determining the validity of the consistency relation.

Assuming that the scalar contribution dominates over the tensor on large scales, i.e.  $r \ll 1$ , one can actually give a measure of the amplitude of the scalar metric perturbation from the observations of the Sachs-Wolfe plateau in the angular power spectrum [21],

$$\left[ \frac{l(l+1)C_l^{(S)}}{2\pi} \right]^{1/2} = \frac{A_S}{5} = (0.926 \pm 0.0106) \times 10^{-5}, \quad (235)$$

$$n = 0.9667 \pm 0.0040, \quad (236)$$

$$\frac{dn}{d \ln k} = -0.002 \pm 0.013. \quad (237)$$

These measurements can be used to normalize the primordial spectrum and determine the parameters of the model of inflation [85]. In the near future these parameters will be determined with much better accuracy, as described in Section 4.4.5.

### 4.4.3 The acoustic peaks

The Sachs-Wolfe plateau is a distinctive feature of Fig. 24. These observations confirm the existence of a primordial spectrum of scalar (density) perturbations on all scales, otherwise the power spectrum would have started from zero at  $l = 2$ . However, we see that the spectrum starts to rise around  $l = 20$  towards the first acoustic peak, where the SW approximation breaks down and the above formulae are no longer valid.

As mentioned above, the first peak in the photon distribution corresponds to overdensities that have undergone half an oscillation, that is, a compression, and appear at a scale associated with the size of the horizon at last scattering, about  $1^\circ$  projected in the sky today. Since photons scatter off baryons, they will also feel the acoustic wave and create a peak in the correlation function. The height of the peak is proportional to the amount of baryons: the larger the baryon content of the universe, the higher the peak. The position of the peak in the power spectrum depends on the geometrical size of the particle horizon at last scattering. Since photons travel along geodesics, the projected size of the causal horizon at decoupling depends on whether the universe is flat, open or closed. In a flat universe the geodesics are straight lines and, by looking at the angular scale of the first acoustic peak, we would be measuring the actual size of the horizon at last scattering. In an open universe, the geodesics are inward-curved trajectories, and therefore the projected size on the sky appears smaller. In this case, the first acoustic peak should occur at higher multipoles or smaller angular scales. On the other hand, for a closed universe, the first peak occurs at smaller multipoles or larger angular scales. The dependence of the position of the first acoustic peak on the spatial curvature can be approximately given by  $l_{\text{peak}} \simeq 220 \Omega_0^{-1/2}$ , where  $\Omega_0 = \Omega_M + \Omega_\Lambda = 1 - \Omega_K$ . Past observations from the balloon experiment BOOMERANG [91], suggested clearly a few years ago that the first peak was between  $l = 180$  and  $250$  at 95% c.l., with an amplitude  $\delta T = 80 \pm 10 \mu\text{K}$ , and therefore the universe was most probably flat. However, with the high precision Planck data we can now pinpoint the spatial curvature to less than a tenth of a percent,

$$\Omega_0 = 0.9992 \pm 0.0040 \quad (95\% \text{ c.l.}) \quad (238)$$

Therefore, the universe is spatially flat (i.e. Euclidean), within 0.1% uncertainty, which is much better than we could ever do before, and is one the most robust predictions of inflation.

With WMAP and specially with Planck, we have now evidence of at least nine distinct acoustic peaks. These peaks should occur at harmonics of the first one, but are typically much lower because of Silk damping. Since the amplitude and position of the primary and secondary peaks are directly determined by the sound speed (and, hence, the equation of state) and by the geometry and expansion of the universe, they can be used as a powerful test of the density of baryons and dark matter, and other cosmological parameters. By looking at these patterns in the anisotropies of the microwave background, cosmologists can determine not only the cosmological parameters, but also the primordial spectrum of density perturbations produced during inflation. It turns out that the observed temperature anisotropies are compatible with a scale-invariant spectrum, see Eq. (236), as predicted by inflation. This is remarkable, and gives very strong support to the idea that inflation may indeed be responsible for both the CMB anisotropies and the large-scale structure of the universe. Different models of inflation have different specific predictions for the fine details associated with the spectrum generated during inflation. It is these minute differences that will allow cosmologists to differentiate between alternative models of inflation and discard those that do not agree with observations. However, most importantly, perhaps, the pattern of anisotropies predicted by inflation is completely different from those predicted by alternative models of structure formation, like cosmic defects: strings, vortices, textures, etc. These are complicated networks of energy density concentrations left over from an early universe phase transition, analogous to the defects formed in the laboratory in certain kinds of liquid crystals when they go through a phase transition. The cosmological defects have spectral properties very different from those generated by inflation. That is why it is so important to launch more sensitive instruments, and with better angular resolution, to determine the properties of the CMB anisotropies.

#### 4.4.4 *The new microwave anisotropy satellites, WMAP and Planck*

The large amount of information encoded in the anisotropies of the microwave background is the reason why both NASA and the European Space Agency have decided to launch two independent satellites to measure the CMB temperature and polarization anisotropies to unprecedented accuracy. The Wilkinson Microwave Anisotropy Probe [92] was launched by NASA at the end of 2000, while Planck [93] was launched by ESA in 2009 and both have fulfilled all our expectations for temperature and E mode polarization. There are at the moment other large mission proposals like PRISM [99], and CORE+ [100], which should provide precision measurements of CMB polarization anisotropies and detect for the first time the primordial B modes of inflation.

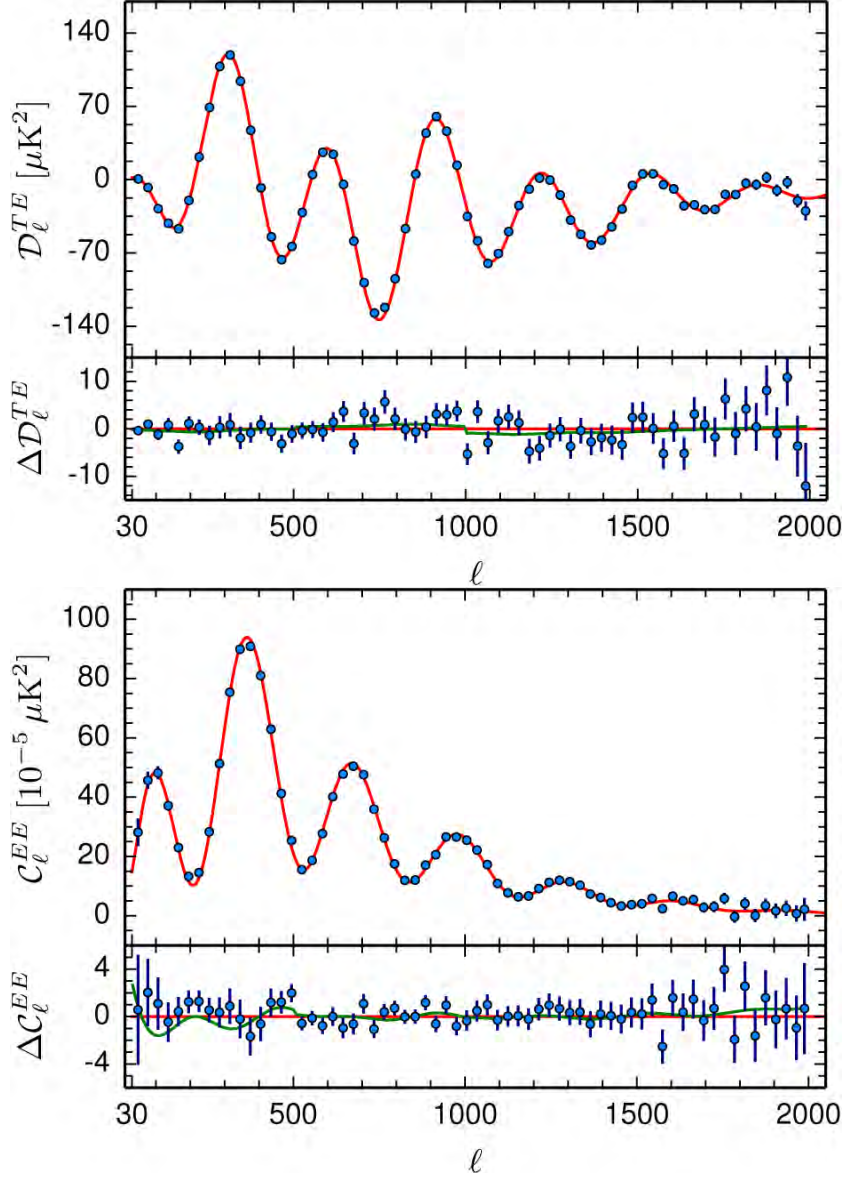
As we have emphasized before, the fact that these anisotropies have such a small amplitude allow for an accurate calculation of the predicted anisotropies in linear perturbation theory. A particular cosmological model is characterized by a dozen or so parameters: the rate of expansion, the spatial curvature, the baryon content, the cold dark matter and neutrino contribution, the cosmological constant (vacuum energy), the reionization parameter (optical depth to the last scattering surface), and various primordial spectrum parameters like the amplitude and tilt of the adiabatic and isocurvature spectra, the amount of gravitational waves, non-Gaussian effects, etc. All these parameters can now be fed into very fast CMB codes called CMBFAST [97] and CAMB [98], that compute the predicted temperature and polarization anisotropies to better than 0.1% accuracy, and thus can be used to compare with observations.

These two satellites have improved both the sensitivity, down to  $\mu\text{K}$ , and the resolution, down to arc minutes, with respect to the previous COBE satellite, thanks to large numbers of microwave horns of various sizes, positioned at specific angles, and also thanks to recent advances in detector technology, with high electron mobility transistor amplifiers (HEMTs) for frequencies below 100 GHz and bolometers for higher frequencies. The primary advantage of HEMTs is their ease of use and speed, with a typical sensitivity of  $0.5 \text{ mKs}^{1/2}$ , while the advantage of bolometers is their tremendous sensitivity, better than  $0.1 \text{ mKs}^{1/2}$ , see Ref. [101]. This has allowed cosmologists to extract information from around 3000 multipoles! Since most of the cosmological parameters have specific signatures in the height and position of the first few acoustic peaks, the higher the resolution, the more peaks one is expected to see, and thus the better the accuracy with which one will be able to measure those parameters, see Table 2.

Although the satellite probes were designed for the accurate measurement of the CMB temperature anisotropies, there are other experiments, like balloon-borne and ground interferometers [94]. Probably the most important objective of the future satellites (beyond WMAP and Planck) will be the measurement of the CMB polarization anisotropies, discovered by DASI in November 2002 [102], and confirmed a few months later by WMAP [21], and by Planck [93] with much greater accuracy, see Fig. 27. These anisotropies were predicted by models of structure formation and indeed found at the level of microKelvin sensitivities, where the new satellites were aiming at. The complementary information contained in the polarization anisotropies already provides much more stringent constraints on the cosmological parameters than from the temperature anisotropies alone. However, in the future, PRISM and CORE+ will have much better sensitivities. In particular, the curl-curl component of the polarization power spectra is nowadays the only means we have to determine the tensor (gravitational wave) contribution to the metric perturbations responsible for temperature anisotropies. If such a component is found, one could finally confirm the inflationary paradigm [95].

#### 4.5 **From metric perturbations to large scale structure**

If inflation is responsible for the metric perturbations that gave rise to the temperature anisotropies observed in the microwave background, then the primordial spectrum of density inhomogeneities induced by the same metric perturbations should also be responsible for the present large scale structure [104]. This simple connection allows for more stringent tests on the inflationary paradigm for the generation of metric perturbations, since it relates the large scales (of order the present horizon) with the smallest scales (on galaxy scales). This provides a very large lever arm for the determination of primordial spectra



**Fig. 27:** Planck measurements of the TE cross-correlation and EE power spectrum of CMB polarization fluctuations, and residuals w.r.t. a Standard Cosmological Model. We only have upper bounds on the BB power spectrum. From Ref. [93].

parameters like the tilt, the nature of the perturbations, whether adiabatic or isocurvature, the geometry of the universe, as well as its matter and energy content, whether CDM, HDM or mixed CHDM.

#### 4.5.1 The galaxy power spectrum

As metric perturbations enter the causal horizon during the radiation or matter era, they create density fluctuations via gravitational attraction of the potential wells. The density contrast  $\delta$  can be deduced from the Einstein equations in linear perturbation theory, see Eq. (165),

$$\delta_k \equiv \frac{\delta\rho_k}{\rho} = \left(\frac{k}{aH}\right)^2 \frac{2}{3} \Phi_k = \left(\frac{k}{aH}\right)^2 \frac{2+2\omega}{5+3\omega} \mathcal{R}_k, \quad (239)$$

where we have assumed  $K = 0$ , and used Eq. (197). From this expression one can compute the power spectrum, at horizon crossing, of matter density perturbations induced by inflation, see Eq. (209),

$$P(k) = \langle |\delta_k|^2 \rangle = A \left( \frac{k}{aH} \right)^n, \quad (240)$$

with  $n$  given by the scalar tilt (211),  $n = 1 + 2\eta - 6\epsilon$ . This spectrum reduces to a Harrison-Zel'dovich spectrum (100) in the slow-roll approximation:  $\eta, \epsilon \ll 1$ .

Since perturbations evolve after entering the horizon, the power spectrum will not remain constant. For scales entering the horizon well after matter domination ( $k^{-1} \gg k_{\text{eq}}^{-1} \simeq 81 \text{ Mpc}$ ), the metric perturbation has not changed significantly, so that  $\mathcal{R}_k(\text{final}) = \mathcal{R}_k(\text{initial})$ . Then Eq. (239) determines the final density contrast in terms of the initial one. On smaller scales, there is a linear transfer function  $T(k)$ , which may be defined as [80]

$$\mathcal{R}_k(\text{final}) = T(k) \mathcal{R}_k(\text{initial}). \quad (241)$$

To calculate the transfer function one has to specify the initial condition with the relative abundance of photons, neutrinos, baryons and cold dark matter long before horizon crossing. The most natural condition is that the abundances of all particle species are uniform on comoving hypersurfaces (with constant total energy density). This is called the *adiabatic* condition, because entropy is conserved independently for each particle species  $X$ , i.e.  $\delta\rho_X = \dot{\rho}_X \delta t$ , given a perturbation in time from a comoving hypersurface, so

$$\frac{\delta\rho_X}{\rho_X + p_X} = \frac{\delta\rho_Y}{\rho_Y + p_Y}, \quad (242)$$

where we have used the energy conservation equation for each species,  $\dot{\rho}_X = -3H(\rho_X + p_X)$ , valid to first order in perturbations. It follows that each species of radiation has a common density contrast  $\delta_r$ , and each species of matter has also a common density contrast  $\delta_m$ , with the relation  $\delta_m = \frac{3}{4}\delta_r$ .

Given the adiabatic condition, the transfer function is determined by the physical processes occurring between horizon entry and matter domination. If the radiation behaves like a perfect fluid, its density perturbation oscillates during this era, with decreasing amplitude. The matter density contrast living in this background does not grow appreciably before matter domination because it has negligible self-gravity. The transfer function is therefore given roughly by, see Eq. (102),

$$T(k) = \begin{cases} 1, & k \ll k_{\text{eq}} \\ (k/k_{\text{eq}})^2, & k \gg k_{\text{eq}} \end{cases} \quad (243)$$

The perfect fluid description of the radiation is far from being correct after horizon entry, because roughly half of the radiation consists of neutrinos whose perturbation rapidly disappears through free streaming. The photons are also not a perfect fluid because they diffuse significantly, for scales below the Silk scale,  $k_S^{-1} \sim 1 \text{ Mpc}$ . One might then consider the opposite assumption, that the radiation has zero perturbation after horizon entry. Then the matter density perturbation evolves according to

$$\ddot{\delta}_k + 2H\dot{\delta}_k + (c_s^2 k_{\text{ph}}^2 - 4\pi G\rho) \delta_k = 0, \quad (244)$$

which corresponds to the equation of a damped harmonic oscillator. The zero-frequency oscillator defines the Jeans wavenumber,  $k_J = \sqrt{4\pi G\rho/c_s^2}$ . For  $k \ll k_J$ ,  $\delta_k$  grows exponentially on the dynamical timescale,  $\tau_{\text{dyn}} = \text{Im } \omega^{-1} = (4\pi G\rho)^{-1/2} = \tau_{\text{grav}}$ , which is the time scale for gravitational collapse. One can also define the Jeans length,

$$\lambda_J = \frac{2\pi}{k_J} = c_s \sqrt{\frac{\pi}{G\rho}}, \quad (245)$$

which separates gravitationally stable from unstable modes. If we define the pressure response timescale as the size of the perturbation over the sound speed,  $\tau_{\text{pres}} \sim \lambda/c_s$ , then, if  $\tau_{\text{pres}} > \tau_{\text{grav}}$ , gravitational collapse of a perturbation can occur before pressure forces can respond to restore hydrostatic equilibrium (this occurs for  $\lambda > \lambda_J$ ). On the other hand, if  $\tau_{\text{pres}} < \tau_{\text{grav}}$ , radiation pressure prevents gravitational collapse and there are damped acoustic oscillations (for  $\lambda < \lambda_J$ ).

We will consider now the behaviour of modes within the horizon during the transition from the radiation ( $c_s^2 = 1/3$ ) to the matter era ( $c_s^2 = 0$ ). The growing mode solution increases only by a factor of 2 between horizon entry and the epoch when matter starts to dominate, i.e.  $y = 1$ . The transfer function is therefore again roughly given by Eq. (243). Since the radiation consists roughly half of neutrinos, which free stream, and half of photons, which either form a perfect fluid or just diffuse, neither the perfect fluid nor the free-streaming approximation looks very sensible. A more precise calculation is needed, including: neutrino free streaming around the epoch of horizon entry; the diffusion of photons around the same time, for scales below Silk scale; the diffusion of baryons along with the photons, and the establishment after matter domination of a common matter density contrast, as the baryons fall into the potential wells of cold dark matter. All these effects apply separately, to first order in the perturbations, to each Fourier component, so that a linear transfer function is produced. There are several parametrizations in the literature, but the one which is more widely used is that of Ref. [105],

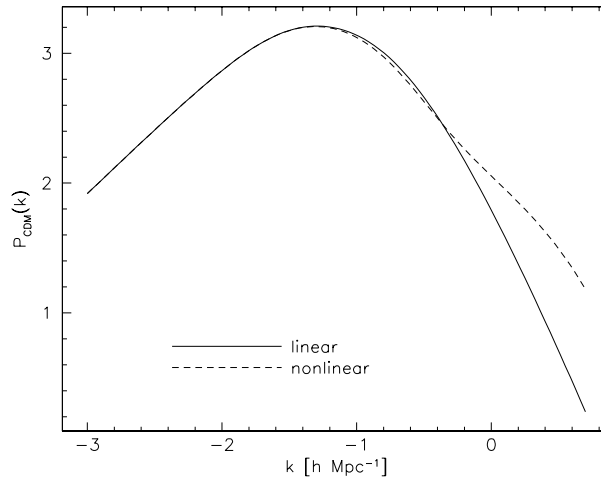
$$T(k) = \left[ 1 + \left( ak + (bk)^{3/2} + (ck)^2 \right)^\nu \right]^{-1/\nu}, \quad \nu = 1.13, \quad (246)$$

$$a = 6.4 (\Omega_M h)^{-1} h^{-1} \text{ Mpc}, \quad (247)$$

$$b = 3.0 (\Omega_M h)^{-1} h^{-1} \text{ Mpc}, \quad (248)$$

$$c = 1.7 (\Omega_M h)^{-1} h^{-1} \text{ Mpc}. \quad (249)$$

We see that the behaviour estimated in Eq. (243) is roughly correct, although the break at  $k = k_{\text{eq}}$  is not at all sharp, see Fig. 28. The transfer function, which encodes the solution to linear equations, ceases to be valid when the density contrast becomes of order 1. After that, the highly nonlinear phenomenon of gravitational collapse takes place, see Fig. 28.



**Fig. 28:** The CDM power spectrum  $P(k)$  as a function of wavenumber  $k$ , in logarithmic scale, normalized to the local abundance of galaxy clusters, for an Einstein-de Sitter universe with  $h = 0.5$ . The solid (dashed) curve shows the linear (non-linear) power spectrum. While the linear power spectrum falls off like  $k^{-3}$ , the non-linear power-spectrum illustrates the increased power on small scales due to non-linear effects, at the expense of the large-scale structures. From Ref. [45].

#### 4.5.2 The new redshift catalogs, 2dF and Sloan Digital Sky Survey

Our view of the large-scale distribution of luminous objects in the universe has changed dramatically during the last 25 years: from the simple pre-1975 picture of a distribution of field and cluster galaxies, to the discovery of the first single superstructures and voids, to the most recent results showing an almost regular web-like network of interconnected clusters, filaments and walls, separating huge nearly empty volumes. The increased efficiency of redshift surveys, made possible by the development of spectrographs and – specially in the last decade – by an enormous increase in multiplexing gain (i.e. the ability to collect spectra of several galaxies at once, thanks to fibre-optic spectrographs), has allowed us not only to do *cartography* of the nearby universe, but also to statistically characterize some of its properties, see Ref. [106]. At the same time, advances in theoretical modeling of the development of structure, with large high-resolution gravitational simulations coupled to a deeper yet limited understanding of how to form galaxies within the dark matter halos, have provided a more realistic connection of the models to the observable quantities [107]. Despite the large uncertainties that still exist, this has transformed the study of cosmology and large-scale structure into a truly quantitative science, where theory and observations can progress side by side.

I will concentrate on two of the new catalogs, which are taking data at the moment and which have changed the field, the 2-degree-Field (2dF) Catalog and the Sloan Digital Sky Survey (SDSS). The advantages of multi-object fibre spectroscopy have been pushed to the extreme with the construction of the 2dF spectrograph for the prime focus of the Anglo-Australian Telescope [46]. This instrument is able to accommodate 400 automatically positioned fibres over a 2 degree in diameter field. This implies a density of fibres on the sky of approximately  $130 \text{ deg}^{-2}$ , and an optimal match to the galaxy counts for a magnitude  $b_J \simeq 19.5$ , similar to that of previous surveys like the ESP, with the difference that with such an area yield, the same number of redshifts as in the ESP survey can be collected in about 10 exposures, or slightly more than one night of telescope time with typical 1 hour exposures. This is the basis of the 2dF galaxy redshift survey. Its goal is to measure redshifts for more than 250,000 galaxies with  $b_J < 19.5$ . In addition, a faint redshift survey of 10,000 galaxies brighter than  $R = 21$  will be done over selected fields within the two main strips of the South and North Galactic Caps. The survey has now finished, with a quarter of a million redshifts. The final result can be seen in Ref. [46].

The most ambitious and comprehensive galaxy survey currently in progress is without any doubt the Sloan Digital Sky Survey [47]. The aim of the project is, first of all, to observe photometrically the whole Northern Galactic Cap,  $30^\circ$  away from the galactic plane (about  $10^4 \text{ deg}^2$ ) in five bands, at limiting magnitudes from 20.8 to 23.3. The expectation is to detect around 50 million galaxies and around  $10^8$  star-like sources. This has already led to the discovery of several high-redshift ( $z > 4$ ) quasars, including the highest-redshift quasar known, at  $z = 5.0$ , see Ref. [47]. Using two fibre spectrographs carrying 320 fibres each, the spectroscopic part of the survey will then collect spectra from about  $10^6$  galaxies with  $r' < 18$  and  $10^5$  AGNs with  $r' < 19$ . It will also select a sample of about  $10^5$  red luminous galaxies with  $r' < 19.5$ , which will be observed spectroscopically, providing a nearly volume-limited sample of early-type galaxies with a median redshift of  $z \simeq 0.5$ , that will be extremely valuable to study the evolution of clustering. The data that is coming from these catalogs is so outstanding that already cosmologists are using them for the determination of the cosmological parameters of the standard model of cosmology. The main outcome of these catalogs is the linear power spectrum of matter fluctuations that give rise to galaxies, and clusters of galaxies. It covers from the large scales of order Gigaparsecs, the realm of the unvirialised superclusters, to the small scales of hundreds of kiloparsecs, where the Lyman- $\alpha$  systems can help reconstruct the linear power spectrum, since they are less sensitive to the nonlinear growth of perturbations.

As often happens in particle physics, not always are observations from a single experiment sufficient to isolate and determine the precise value of the parameters of the standard model. We mentioned in the previous Section that some of the cosmological parameters created similar effects in the temperature anisotropies of the microwave background. We say that these parameters are *degenerate* with

respect to the observations. However, often one finds combinations of various experiments/observations which break the degeneracy, for example by depending on a different combination of parameters. This is precisely the case with the cosmological parameters, as measured by a combination of large-scale structure observations, microwave background anisotropies, Supernovae Ia observations and Hubble Space Telescope measurements. It is expected that in the near future we will be able to determine the parameters of the standard cosmological model with great precision from a combination of several different experiments.

## 5 Conclusions

In the last ten years we have seen a true revolution in the quality and quantity of cosmological data that has allowed cosmologists to determine most of the cosmological parameters with a few percent accuracy and thus fix a Standard Model of Cosmology. The art of measuring the cosmos has developed so rapidly and efficiently that one may be tempted of renaming this science as Cosmonomy, leaving the word Cosmology for the theories of the Early Universe. In summary, we now know that the stuff we are made of – baryons – constitutes just about 4% of all the matter/energy in the Universe, while 25% is dark matter – perhaps a new particle species related to theories beyond the Standard Model of Particle Physics –, and the largest fraction, 70%, some form of diffuse tension also known as dark energy – perhaps a cosmological constant. The rest, about 1%, could be in the form of massive neutrinos.

Nowadays, a host of observations – from CMB anisotropies and large scale structure to the age and the acceleration of the universe – all converge towards these values, see Fig. 25. Fortunately, we will have, within this decade, new satellite experiments like Planck, CMBpol, SNAP as well as deep galaxy catalogs from Earth, to complement and precisely pin down the values of the Standard Model cosmological parameters below the percent level, see Table 2.

All these observations would not make much sense without the encompassing picture of the inflationary paradigm that determines the homogeneous and isotropic background on top of which it imprints an approximately scale invariant gaussian spectrum of adiabatic fluctuations. At present all observations are consistent with the predictions of inflation and hopefully in the near future we may have information, from the polarization anisotropies of the microwave background, about the scale of inflation, and thus about the physics responsible for the early universe dynamics.

## Acknowledgements

I would like to thank the organizers of the CERN Latin American School of High Energy Physics 2013, and very specially Martijn Mulders, for his incredible patience with my contribution. This work was supported in part by a CICYT project FPA2012-39684-C03-02.

## References

- [1] A. Einstein, Sitz. Preuss. Akad. Wiss. Phys. **142** (1917) (§4); Ann. Phys. **69** (1922) 436.
- [2] A. Friedmann, Z. Phys. **10** (1922) 377.
- [3] E.P. Hubble, Publ. Nat. Acad. Sci. **15** (1929) 168.
- [4] G. Gamow, Phys. Rev. **70** (1946) 572; Phys. Rev. **74** (1948) 505.
- [5] A.A. Penzias and R.W. Wilson, Astrophys. J. **142** (1965) 419.
- [6] S. Weinberg, *Gravitation and Cosmology* (John Wiley & Sons, San Francisco, 1972).
- [7] S. Perlmutter *et al.* [Supernova Cosmology Project], Astrophys. J. **517** (1999) 565. Home Page <http://scp.berkeley.edu/>
- [8] W. L. Freedman *et al.*, Astrophys. J. **553** (2001) 47.
- [9] A. G. Riess *et al.* [High-z Supernova Search], Astron. J. **116** (1998) 1009. Home Page <http://cfa-www.harvard.edu/cfa/oir/Research/supernova/>



- [10] J. García-Bellido in *European School of High Energy Physics 2004*, ed. R. Fleischer (CERN report 2006-003); e-print Archive: astro-ph/0502139.
- [11] R. A. Knop *et al.* [Supernova Cosmology Project Collaboration], *Astrophys. J.* **598** (2003) 102 [astro-ph/0309368].
- [12] A. G. Riess *et al.* [Supernova Search Team Collaboration], *Astrophys. J.* **607** (2004) 665 [astro-ph/0402512].
- [13] R. Amanullah, *et al.*, *Astrophys. J.* **716** (2010) 712 [arXiv:1004.1711 [astro-ph.CO]].
- [14] S. Weinberg, *Rev. Mod. Phys.* **61** (1989) 1; S. M. Carroll, *Living Rev. Rel.* **4** (2001) 1; T. Padmanabhan, *Phys. Rept.* **380** (2003) 235; P. J. E. Peebles and B. Ratra, *Rev. Mod. Phys.* **75** (2003) 559.
- [15] The SuperNova/Acceleration Probe Home page: <http://snap.lbl.gov/>
- [16] E.W. Kolb and M.S. Turner, "The Early Universe", Addison Wesley (1990).
- [17] R. Srianand, P. Petitjean and C. Ledoux, *Nature* **408** (2000) 931.
- [18] S. Burles, K.M. Nollett, J.N. Truran, M.S. Turner, *Phys. Rev. Lett.* **82** (1999) 4176; S. Burles, K.M. Nollett, M.S. Turner, "Big-Bang Nucleosynthesis: Linking Inner Space and Outer Space", e-print Archive: astro-ph/9903300.
- [19] K. A. Olive, G. Steigman and T. P. Walker, *Phys. Rept.* **333** (2000) 389; J. P. Kneller and G. Steigman, "BBN For Pedestrians," *New J. Phys.* **6** (2004) 117.
- [20] Particle Data Group Home Page, <http://pdg.web.cern.ch/pdg/>
- [21] D. N. Spergel *et al.*, *Astrophys. J. Suppl.* **148** (2003) 175.
- [22] J.C. Mather *et al.*, *Astrophys. J.* **512** (1999) 511.
- [23] R.H. Dicke, P.J.E. Peebles, P.G. Roll and D.T. Wilkinson, *Astrophys. J.* **142** (1965) 414.
- [24] C.L. Bennett *et al.*, *Astrophys. J.* **464** (1996) L1.
- [25] P.J.E. Peebles, "Principles of Physical Cosmology", Princeton U.P. (1993).
- [26] T. Padmanabhan, "Structure Formation in the Universe", Cambridge U.P. (1993).
- [27] E.R. Harrison, *Phys. Rev. D* **1** (1970) 2726; Ya. B. Zel'dovich, *Astron. Astrophys.* **5** (1970) 84.
- [28] The IRAS Point Source Catalog Web page:  
<http://www-astro.physics.ox.ac.uk/~wjs/pscz.html>
- [29] P.J. Steinhardt, in *Particle and Nuclear Astrophysics and Cosmology in the Next Millennium*, ed. by E.W. Kolb and R. Peccei (World Scientific, Singapore, 1995).
- [30] W.L. Freedman, "Determination of cosmological parameters", Nobel Symposium (1998), e-print Archive: hep-ph/9905222.
- [31] S. Refsdal, *Mon. Not. R. Astr. Soc.* **128** (1964) 295; **132** (1966) 101.
- [32] R.D. Blandford and T. Kundić, "Gravitational Lensing and the Extragalactic Distance Scale", e-print Archive: astro-ph/9611229.
- [33] N.A. Grogin and R. Narayan, *Astrophys. J.* **464** (1996) 92.
- [34] M. Birkinshaw, *Phys. Rep.* **310** (1999) 97.
- [35] The Chandra X-ray observatory Home Page: <http://chandra.harvard.edu/>
- [36] W. L. Freedman *et al.*, *Astrophys. J.* **553** (2001) 47
- [37] F. Zwicky, *Helv. Phys. Acta* **6** (1933) 110.
- [38] H. Babcock, *Pub. Astr. Soc. Pacific* **50** (1938) 174.
- [39] K. C. Freeman, *Astrophys. J.* **160** (1970) 811.
- [40] V. C. Rubin, *Science* **209** (1980) 63; V. C. Rubin, N. Thonnard and W. K. Ford, *Astrophys. J.* **238** (1980) 471.
- [41] M. Persic, P. Salucci and F. Stel, *Mon. Not. Roy. Astron. Soc.* **281** (1996) 27.

- [42] C.M. Baugh *et al.*, “Ab initio galaxy formation”, e-print Archive: astro-ph/9907056; *Astrophys. J.* **498** (1998) 405.
- [43] F. Prada *et al.*, *Astrophys. J.* **598** (2003) 260.
- [44] C.L. Sarazin, *Rev. Mod. Phys.* **58** (1986) 1.
- [45] M. Bartelmann *et al.*, *Astron. & Astrophys.* **330** (1998) 1; M. Bartelmann and P. Schneider, *Phys. Rept.* **340** (2001) 291
- [46] M. Colless *et al.* [2dFGRS Coll.], “The 2dF Galaxy Redshift Survey: Final Data Release,” Archive: astro-ph/0306581. The 2dFGRS Home Page: <http://www.mso.anu.edu.au/2dFGRS/>
- [47] M. Tegmark *et al.* [SDSS Collaboration], *Astrophys. J.* **606** (2004) 702; *Phys. Rev. D* **69** (2004) 103501. The SDSS Home Page: <http://www.sdss.org/sdss.html>
- [48] G.G. Raffelt, “Dark Matter: Motivation, Candidates and Searches”, European Summer School of High Energy Physics 1997. CERN Report pp. 235-278, e-print Archive: hep-ph/9712538.
- [49] P.J.E. Peebles, “Testing GR on the Scales of Cosmology,” e-print Archive: astro-ph/0410284.
- [50] M. C. Gonzalez-Garcia, “Global analysis of neutrino data,” e-print Archive: hep-ph/0410030.
- [51] S.D. Tremaine and J.E. Gunn, *Phys. Rev. Lett.* **42** (1979) 407; J. Madsen, *Phys. Rev. D* **44** (1991) 999.
- [52] J. Primack, D. Seckel and B. Sadoulet, *Ann. Rev. Nucl. Part. Sci.* **38** (1988) 751; N.E. Booth, B. Cabrera and E. Fiorini, *Ann. Rev. Nucl. Part. Sci.* **46** (1996) 471.
- [53] C. Kraus *et al.*, “Final results from phase II of the Mainz neutrino mass search in tritium beta decay,” e-print Archive: hep-ex/0412056.
- [54] H. V. Klapdor-Kleingrothaus *et al.*, *Mod. Phys. Lett. A* **16** (2001) 2409; *Mod. Phys. Lett. A* **18** (2003) 2243.
- [55] R. Bernabei, *et al.*, *Int. J. Mod. Phys. A* **28** (2013) 1330022  
R. Bernabei *et al.*, “Dark matter search,” *Riv. Nuovo Cim.* **26N1** (2003) 1. DAMA Home Page, <http://www.lngs.infn.it/lngs/htxts/dama/welcome.html>
- [56] K. A. Olive, “Dark matter candidates in supersymmetric models,” e-print Archive: hep-ph/0412054.
- [57] R. Agnese *et al.* [SuperCDMS Collaboration], *Phys. Rev. Lett.* **112** (2014) 24, 241302.
- [58] B. Ahmed *et al.*, *Nucl. Phys. Proc. Suppl.* **124** (2003) 193; *Astropart. Phys.* **19** (2003) 691. UKDMC Home Page at <http://hepwww.rl.ac.uk/ukdmc/>
- [59] M. Bravin *et al.*, *Astropart. Phys.* **12** (1999) 107.
- [60] J. I. Collar *et al.*, *Phys. Rev. Lett.* **85** (2000) 3083.
- [61] G. Jungman, M. Kamionkowski and K. Griest, *Phys. Rep.* **267** (1996) 195.
- [62] The Alpha Magnetic Spectrometer Home Page: <http://ams.cern.ch/AMS/>
- [63] F. Halzen *et al.*, *Phys. Rep.* **307** (1998) 243.
- [64] D.A. Vandenberg, M. Bolte and P.B. Stetson, *Ann. Rev. Astron. Astrophys.* **34** (1996) 461; e-print Archive: astro-ph/9605064.
- [65] L. M. Krauss, *Phys. Rept.* **333** (2000) 33.
- [66] B. Chaboyer, P. Demarque, P.J. Kernan and L.M. Krauss, *Science* **271** (1996) 957; *Astrophys. J.* **494** (1998) 96.
- [67] C.H. Lineweaver, *Science* **284** (1999) 1503.
- [68] D. Scott, J. Silk and M. White, *Science* **268** (1995) 829; W. Hu, N. Sugiyama and J. Silk, *Nature* **386** (1997) 37; E. Gawiser and J. Silk, *Phys. Rept.* **333** (2000) 245.
- [69] J. Silk, *Nature* **215** (1967) 1155.
- [70] A. Albrecht, R. A. Battye and J. Robinson, *Phys. Rev. Lett.* **79** (1997) 4736; N. Turok, U. L. Pen, U. Seljak, *Phys. Rev. D* **58** (1998) 023506; L. Pogosian, *Int. J. Mod. Phys. A* **16S1C** (2001) 1043.

- [71] U. Seljak et al., [SDSS Collaboration], *Phys. Rev. D* **71** (2005) 103515.
- [72] V. Barger, D. Marfatia and A. Tregre, *Phys. Lett. B* **595** (2004) 55; P. Crotty, J. Lesgourgues and S. Pastor, *Phys. Rev. D* **69** (2004) 123007; S. Hannestad, *Nucl. Phys. Proc. Suppl.* **145** (2005) 313.
- [73] H. V. Peiris *et al.*, *Astrophys. J. Suppl.* **148** (2003) 213; P. Crotty, J. García-Bellido, J. Lesgourgues and A. Riazuelo, *Phys. Rev. Lett.* **91** (2003) 171301; J. Valiviita and V. Muhonen, *Phys. Rev. Lett.* **91** (2003) 131302; M. Beltrán, J. García-Bellido, J. Lesgourgues and A. Riazuelo, *Phys. Rev. D* **70** (2004) 103530; K. Moodley, M. Bucher, J. Dunkley, P. G. Ferreira and C. Skordis, *Phys. Rev. D* **70** (2004) 103520; C. Gordon and K. A. Malik, *Phys. Rev. D* **69** (2004) 063508; F. Ferrer, S. Rasanen and J. Valiviita, *JCAP* **0410** (2004) 010; H. Kurki-Suonio, V. Muhonen and J. Valiviita, *Phys. Rev. D* **71** (2005) 063005; M. Beltrán, J. García-Bellido, J. Lesgourgues, A. R. Liddle and A. Slosar, *Phys. Rev. D* **71** (2005) 063532; M. Beltran, J. Garcia-Bellido, J. Lesgourgues and M. Viel, *Phys. Rev. D* **72** (2005) 103515; M. Beltran, J. Garcia-Bellido and J. Lesgourgues, *Phys. Rev. D* **75** (2007) 103507.
- [74] J. García-Bellido, *Phil. Trans. R. Soc. Lond. A* **357** (1999) 3237.
- [75] A. Guth, *Phys. Rev. D* **23** (1981) 347.
- [76] A.D. Linde, *Phys. Lett.* **108B** (1982) 389.
- [77] A. Albrecht and P.J. Steinhardt, *Phys. Rev. Lett.* **48** (1982) 1220.
- [78] For a personal historical account, see A. Guth, “The Inflationary Universe”, Perseus Books (1997).
- [79] A.D. Linde, “Particle Physics and Inflationary Cosmology”, Harwood Academic Press (1990).
- [80] A.R. Liddle and D.H. Lyth, *Phys. Rep.* **231** (1993) 1.
- [81] J.M. Bardeen, *Phys. Rev. D* **22** (1980) 1882.
- [82] V.F. Mukhanov, H.A. Feldman and R.H. Brandenberger, *Phys. Rep.* **215** (1992) 203.
- [83] M. Abramowitz and I. Stegun, “Handbook of Mathematical Functions”, Dover (1972).
- [84] J. García-Bellido and D. Wands, *Phys. Rev. D* **53** (1996) 5437; D. Wands, K. A. Malik, D. H. Lyth and A. R. Liddle, *Phys. Rev. D* **62** (2000) 043527.
- [85] D.H. Lyth and A. Riotto, *Phys. Rep.* **314** (1999) 1.
- [86] R.K. Sachs and A.M. Wolfe, *Astrophys. J.* **147** (1967) 73.
- [87] E.R. Harrison, *Rev. Mod. Phys.* **39** (1967) 862; L.F. Abbott and R.K. Schaefer, *Astrophys. J.* **308** (1986) 546.
- [88] E.F. Bunn, A.R. Liddle and M. White, *Phys. Rev. D* **54** (1996) 5917.
- [89] A.A. Starobinsky, *Sov. Astron. Lett.* **11** (1985) 133.
- [90] S.M. Carroll, W.H. Press and E.L. Turner, *Ann. Rev. Astron. Astrophys.* **30** (1992) 499.
- [91] P. de Bernardis *et al.*, *New Astron. Rev.* **43** (1999) 289; P. D. Mauskopf *et al.* [Boomerang Coll.], *Astrophys. J.* **536** (2000) L59. Boomerang Home Page: <http://oberon.roma1.infn.it/boomerang/>
- [92] Microwave Anisotropy Probe Home Page: <http://map.gsfc.nasa.gov/>
- [93] Planck Surveyor Home Page: <http://www.cosmos.esa.int/web/planck/publications>
- [94] M. Tegmark Home Page: <http://space.mit.edu/home/tegmark/>
- [95] M. Kamionkowski and A. Kosowsky, *Ann. Rev. Nucl. Part. Sci.* **49** (1999) 77.
- [96] W. Hu and S. Dodelson, *Ann. Rev. Astron. Astrophys.* **40** (2002) 171.
- [97] U. Seljak and M. Zaldarriaga, CMBFAST code Home Page: <http://www.cmbfast.org/>
- [98] A. Lewis and A. Challinor, CAMB code Home Page: <http://camb.info/>
- [99] P. Andre *et al.* [PRISM Collaboration], arXiv:1306.2259 [astro-ph.CO]; *JCAP* **1402** (2014) 006 [arXiv:1310.1554 [astro-ph.CO]].  
PRISM experiment Home Page: <http://www.prism-mission.org/>
- [100] F. R. Bouchet *et al.* [COre Collaboration], “COre (Cosmic Origins Explorer) A White Paper,” arXiv:1102.2181 [astro-ph.CO]; CORE experiment Home Page:

<http://www.core-mission.org/>

- [101] L.A. Page, “Measuring the anisotropy in the CMB”, e-print Archive: astro-ph/9911199.
- [102] J. Kovac et al., Nature **420** (2002) 772; DASI Home Page: <http://astro.uchicago.edu/dasi/>
- [103] BICEP & Keck Array Home Page, <http://bicepkeck.org/>
- [104] A.R. Liddle and D.H. Lyth, “Cosmological Inflation and Large Scale Structure”, Cambridge University Press (2000).
- [105] J.R. Bond and G. Efstathiou, Astrophys. J. **285** (1984) L45.
- [106] G. Efstathiou *et al.* (Eds.) “Large-scale structure in the universe”, Phil. Trans. R. Soc. Lond. **A 357** (1999) 1-198.
- [107] B. Moore, Phil. Trans. R. Soc. Lond. **A 357** (1999) 3259.

## Organizing Committee

M. Aguilar (CIEMAT, Spain)  
L. Álvarez-Gaumé (CERN)  
F. Barrio (CIEMAT, Spain)  
C. Dib (UTFSM, Chile)  
M.T. Dova (UNLP, Argentina)  
J. Ellis (King's College London, UK and CERN)  
N. Ellis (Schools Director, CERN (Chair))  
A. Gago (PUCP, Peru)  
M. Gandelman (UFRJ, Brazil)  
P. Garcia (CIEMAT, Spain)  
C. Grojean (CERN and IFAE-ICREA, Barcelona, Spain)  
H. Haller (Schools Administrator, CERN)  
M. Losada (UAN, Colombia)  
M. Mulders (Schools Deputy-Director, CERN)  
G. Perez (CERN and Weizmann Institute, Israel)  
K. Ross (Schools Administrator, CERN)  
A. Zepeda (Cinvestav, Mexico)

## Local Organizing Committee

H. Castillo (PUCP, Peru)  
A. Gago (PUCP, Peru (Chair))  
D. Pacheco (UNSA, Peru)  
R. Perca (UNSA, Peru)  
O. Pereyra (UNI, Peru)  
J. Solano (UNI, Peru)

## Lecturers

L. Álvarez-Gaumé (CERN)  
F. Maltoni (UC Louvain-CP3, Belgium)  
Y. Nir (Weizmann Institute, Israel)  
C. Gonzalez-Garcia (Stony Brook, USA and U. Barcelona, Spain)  
G. Burdman (USP, Brazil)  
E.S. Fraga (UFRJ, Brazil)  
W. Riegler (CERN)  
O. González (CIEMAT, Spain)  
K. Cranmer (NYU, USA)  
J. García-Bellido (IFT Madrid, Spain)  
M.T. Dova (UNLP and CONICET, Argentina)

## Discussion Leaders

D. Delepine (Guanajuato U., Mexico)  
L. Diaz Cruz (BUAP, Mexico)  
J. Jones (PUCP, Peru)  
A. Zerwekh (UTFSM, Chile)

## Students

Thamys ABRAHÃO  
María Josefina ALCONADA VERZINI  
Francisco ALONSO  
Pedro AMAO  
Carmen ARAUJO DEL CASTILLO  
Alexander ARGUELLO QUIROGA  
María José BUSTAMANTE ROSELL  
Juan Enrique CALDERON KREJCI  
Anthonny Freddy CANAZAS GARAY  
Jose Alonso CARPIO DUMLER  
Daniela del Rocio CARRILLO CURO  
Gabriela CERQUEIRA GOMES  
Alan Gilberto CHÁVEZ MEZA  
Estefania COLUCCIO LESKOW  
Frank CORONADO  
Giovanna COTTIN  
Percy CÁCERES  
Alexandru DAFINCA  
Orjan DALE  
Bruno DANIEL  
Lucas DE BRITO CAVALCANTI  
Alejandro DE LA PUENTE  
Margot DELGADO DE LA FLOR CASTRO  
Samantha DOOLING  
Gonzalo DÍAZ BAUTISTA  
Miguel Francisco GARCIA VERA  
Jose GARCIA  
Alejandro GOMEZ  
Marcela GONZÁLEZ  
Luis Max GUILLEN QUIROZ  
David HALL  
Alberto HERNANDEZ ALMADA  
Maurício HIPPERT TEIXEIRA  
Pablo JACOME  
Cecilia Gisele JARNE  
José Carlos JIMÉNEZ APAZA  
Matthew KENZIE  
Jose Luis LA ROSA NAVARRO  
Duncan LEGGAT  
Ivonne Alicia MALDONADO CERVANTES  
Pedro MALTA  
Alessandro MANFREDINI  
Rosana MARTINEZ TURTOS  
Sony MARTINS  
Karim MASSRI  
Bernabe Alonso MEJIA CORDERO  
Jhovanny MEJIA  
Andrés MELO  
Nicolás MILEO  
Josué MOLINA  
Gibraham Ivanhoe NAPOLES CANEDO  
Roger Felipe NARANJO GARCIA  
Guillermo PALACIO  
Saul PANIBRA CHURATA  
Alexander PARADA  
Lars Egholm PEDERSEN  
Denis ROBERTSON  
Silvestre ROMANO  
Sabrina SACERDOTI  
Glauber SAMPAIO DOS SANTOS  
Adriano SAMPIERI  
Pierre SAOUTER  
Karoline SELBACH  
Simao Paulo SILVA  
Mauricio SUÁREZ  
Vanessa THEODORO  
Edgar VALENCIA-RODRIGUEZ  
Janeth VALVERDE  
Teofilo VARGAS  
oscar VARGAS  
Juan Pablo VELÁSQUEZ ORMAECHE  
Cynthia VIZCARRA VENTURA  
Ivan Francisco YUPANQUI TELLO  
Jilberto ZAMORA SAA  
Dennis ZAVALETA

## Posters

Author	Poster title
J. ALCONADA	Material Estimation of the ATLAS Inner Detector with photon conversions
F. ALONSO	Search for single photon events with large missing $E_T$ in pp collisions at LHC with the ATLAS detector
A. CHAVEZ, H. FALCÓN, L. VILLASEÑOR	Correlation of ultra-high-energy cosmic rays with nearby active galactic nuclei using distance-dependent and flux-dependent weights
E. COLUCCIO LESKO, E. ALVAREZ	A charged $Z'$ to explain the apparent disagreement in top-antitop asymmetries between Tevatron and the LHC
G. COTTIN	ATLAS Tau Trigger
A. DAFINCA	Search for direct sbottom pair production in events with missing transverse momentum and two b-jets in $12.8 \text{ fb}^{-1}$ of pp collisions at $\sqrt{s} = 8 \text{ TeV}$ with the ATLAS detector
B. DANIEL	Irregular stations on the event reconstruction of the Pierre Auger Observatory
S. DOOLING	Nonperturbative and Parton Shower Corrections in matched NLO-shower event generators
C. JARNE	Risetime at 1000 m: Searching for a new SD observable sensitive to mass composition
D. HALL	Theoretical uncertainties on the WW background to the $H \rightarrow WW$ search with the ATLAS detector
R. MARTINEZ TURTOS, Z. CONESA DEL VALLE	Evaluation of inclusive $J/\Psi$ production cross section, rapidity and $p^T$ distributions for the pPb LHC run at $\sqrt{s} = 5.023 \text{ TeV}$
S. MARTINS	Jet Fragmentation in Heavy Ion Collisions
K. MASSRI	The Kaon Identification Detector for the NA62 experiment at CERN
N. MILEO, K. KIERS, A. SZYNKMAN	Angular correlations in $\tau \rightarrow K\pi\pi\nu$
R. NARANJO	Studies of low $p^T$ photons for the Higgs to Zgamma analysis

<b>Author</b>	<b>Poster title</b>
M. TEIXEIRA, E. FRAGA, E. SANTOS	Boundary Condition Effects in Heavy Ion Collisions
K. SELBACH	Search for the Standard Model Higgs boson decay $H \rightarrow ZZ^{(*)} \rightarrow 4l$ with the ATLAS detector
S.P. SILVA	Determination of the Forward-Backward Asymmetry in $e^+e^-$ pair production via torsion at the LHC
E. VALENCIA	Neutrino-Electron Elastic Scattering in MINERvA at 5 GeV
J. ZAMORA, G. CVETIC	Heavy Majorana Neutrinos and their role in Meson Decay