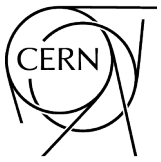
CERN Yellow Reports: School Proceedings
Volume 2/2017

CERN-2017-005-SP
KEK Proceedings 2017-3

2014 Asia–Europe–Pacific School of High-Energy Physics

Puri, India
4 – 17 November 2014

Editors: M. Mulders
R. Godbole



CERN Yellow Reports: School Proceedings
Published by CERN, CH-1211 Geneva 23, Switzerland

ISBN 978–92–9083–460–1 (paperback)
ISBN 978–92–9083–461–8 (PDF)
ISSN 2519-8041 (Print)
ISSN 2519-805X (Online)
DOI https://doi.org/10.23730/CYRSP-2017-002

Accepted for publication by the CERN Report Editorial Board (CREB) on 24 July 2017

Available online at http://publishing.cern.ch/ and http://cds.cern.ch/

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

This volume should be cited as:

Proceedings of the 2014 Asia–Europe–Pacific School of High-Energy Physics, Puri, India, 4 – 17 November 2014, edited by M. Mulders and R. Godbole, CERN Yellow Reports: School Proceedings, Vol. 2/2017, CERN-2017-005-SP (CERN, Geneva, 2017), https://doi.org/10.23730/CYRSP-2017-002

A contribution in this volume should be cited as:

[Author name(s)], in Proceedings of the 2014 Asia–Europe–Pacific School of High-Energy Physics, Puri, India, 4 – 17 November 2014, edited by M. Mulders and R. Godbole, CERN Yellow Reports: School Proceedings, Vol. 2/2017, CERN-2017-005-SP (CERN, Geneva, 2017), pp. [first page]–[last page], https://doi.org/10.23730/CYRSP-2017-002.[first page]

Abstract

The Asia–Europe–Pacific 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 quantum field theory and the electroweak Standard Model, the theory of quantum chromodynamics, flavour physics and CP violation, neutrino physics, heavy-ion physics, cosmology and a brief introduction to the principles of instrumentation and detectors for particle physics.

Preface

The second event in the series of the Asia–Europe–Pacific School of High-Energy Physics took place in Puri, India, from 4 to 17 November 2014. A strong team from IISc, IOP, NISER, TIFR and VECC, took care of the local local organization, while CERN and KEK collaborated to provide administrative support in preparation for the School.

The staff and students were housed in comfortable accommodation in the Toshali Sands Hotel that also provided the conference facilities. The students shared twin-bed rooms, mixing nationalities to foster cultural exchange between participants from different countries.

A total of 64 students of 22 different nationalities attended the school. About 70% of the students were from Asia-Pacific countries, most of the others coming from Europe. More than 80% of the participants were working towards a PhD, while most of the others were advanced Masters students; the School was also open to postdocs. Over 80% of the students were experimentalists; the school was also open to phenomenologists.

A total of 33 lectures were complemented by daily discussion sessions led by five discussion leaders. The teachers (lecturers and discussion leaders) came from many different countries: Australia, China, France, Germany, India, Japan, Korea, the Netherlands, Russia, Switzerland and Taiwan.

The programme required the active participation of the students. In addition to the discussion sessions that addressed questions from the lecture courses, there was an evening session in which many students presented posters about their own research work to their colleagues and the teaching staff.

Collaborative student projects in which the students of each discussion group worked together on an in-depth study of a published experimental data analysis were an important activity. This required interacting, outside of the formal teaching sessions, with colleagues from different countries and different cultures. A student representative of each of the five groups presented a short summary of the conclusions of the group's work in a special evening session.

In addition to the academic side of the School, the participants had the occasion to experience many aspects of Indian culture, including visits to the Sun Temple at Konark, to numerous sites and temples in and around the city of Bhubaneswar, and to observe the natural beauty of Lake Chilka and its associated wildlife. They also had ample opportunity to appreciate excellent Indian food, including some delicious dinners served in the open air with live performances of Indian dance.

Our thanks go to the local-organization team and, in particular, to Subhasis Chattopadhyay, Rohini Godbole, Gobinda Majumdar, Prolay K. Mal, Sreerup Raychaudhuri and Pradip K. Sahu, for all their work and assistance in preparing the School, on both scientific and practical matters, and for their presence throughout the event. Our thanks also go to the hotel management and staff who assisted the School organizers and the participants in many ways.

Very great 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 listening to and discussing with the teaching staff of world renown.

We would like to express our special appreciation to Professor Rolf Heuer, Director General of CERN, and Professor Atsuto Suzuki, Director General of KEK, for their lectures on the particle-physics programmes in Europe and in Asia. We would also like to thank Professor K. Vijayaraghavan, Secretary of the Department of Science and Technology, for his welcome address, and Professor Naba Mondal, leader of the India-based Neutrino Observatory, for his presentation on high-energy physics in India.

We are very grateful to Kate Ross from CERN, and to Misa Miyai and Ritsuko Ota from KEK, for their untiring efforts on administration for the School. We would also like to thank the members of the International Committees

Sponsorship from numerous bodies in many countries covered the cost of travel and/or local expenses of their staff and students who attended the School. In addition, general sponsorship is gratefully acknowledged from: Bose Institute, India; CEA/Irfu, France; CNRS/IN2P3, France; CERN; DESY, Germany; ICTP; KEK, Japan; TIFR, India.

Nick Ellis
(Chair of the International Organizing Committee)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72

People in the photograph

1	Raveendrababu Karanam	37	Aleksandr Azatov
2	Divekar S.T.	38	Lucia Grillo
3	Reza Goldouzian	39	Artur Shaikhiev
4	Ashim Roy	40	Debjyoti Bardhan
5	Kalyanmoy Chatterje	41	SK Noor Alam
6	Martijn Mulders	42	Aliaksei Hrynevich
7	Nick Ellis	43	Valery Rubakov
8	Sandeep Bhowmik	44	Sudhir Vempati
9	Rajesh Ganai	45	Chandan Gupta
10	Rohini Godbole	46	Simon Stark Mortensen
11	Subikash Choudhury	47	Timofey Maltsev
12	Jayita Lahiri	48	Andrey Kupich
13	Alexander Bylinkin	49	Weimin Song
14	Tatsuhiko Tomita	50	Prasanth Krishnan KP
15	Deepanjali Goswami	51	Md. Mohsin
16	Ipsita Saha	52	Ram Krishna Dewarjee
17	Indrani Chakraborty	53	Tom Ravenscroft
18	Norma Sidik Risdianto	54	Anastasiia Kozachuk
19	Nairit Sur	55	Ievgen Korol
20	Subhasis Chattopadhyay	56	Kouhei Hanzawa
21	Koichi Hamaguchi	57	Ralitsa Sharankova
22	Robyn Lucas	58	Alexey Baskakov
23	Sylvestre Pires	59	Niladribihari Sahoo
24	Rose Koopman	60	Inayat Bhat
25	Yusho Homma	61	Bibhuprasad Mahakud
26	Ryutaro Nishimura	62	Dibyakrupa Sahoo
27	Subash Adhikari	63	Fabian Kuger
28	Narayan Rana	64	Genessis Perez
29	Soumita Pramanick	65	Lydia Roos
30	Kuo-Lun Jen	66	Tran Nam
31	Yuki Nakai	67	Monika Blanke
32	Nadine Fischer	68	Ian-Woo Kim
33	Sijing Zhang	69	Soureek Mitra
34	Li-Chu Chang	70	Cheng-Wei Chang
35	Kate Ross	71	Neil Barrie
36	Dinesh Kumar	72	Marian Stahl

PHOTOGRAPHS (MONTAGE)


Atmospheric neutrinos
Cosmic rays entering the atmosphere interact & produce π+
Charged π then decay dominantly as:
Super-Kamiokande
In case of no oscillation
However measurements showed
Is neutrino oscillation the explanation?
AERSHCP
2010

Contents

Proceedings of the 2014 Asia–Europe–Pacific School of High-Energy Physics, Puri, India, 4–17 November 2014, edited by
M. Mulders and R. Godbole, CERN Yellow Reports: School Proceedings, Vol. 2/2017, CERN-2017-005-SP (CERN, Geneva, 2017)

Field Theory and the Electro-Weak Standard Model

R. M. Godbole
Centre for High Energy Physics, Indian Institute of Science, Bangalore, India.

Abstract

In this set of four lectures I will discuss some aspects of the Standard Model (SM) as a quantum field theory and related phenomenological observations which have played a crucial role in establishing the $SU(2)_L \times U(1)_Y$ gauge theory as the correct description of Electro-Weak (EW) interactions. I will first describe in brief the idea of EW unification as well as basic aspects of the Higgs mechanism of spontaneous symmetry breaking. After this I will discuss anomaly cancellation, custodial symmetry and implications of the high energy behavior of scattering amplitudes for the particle spectrum of the EW theory. This will be followed up by a discussion of the 'indirect' constraints on the SM particle masses such as M_c, M_t and M_h from various precision EW measurements. I will end by discussing the theoretical limits on M_h and implications of the observed Higgs mass for the SM and beyond.

Keywords

Lectures; Standard Model; electroweak interaction; gauge theory; spontaneous symmetry breaking; field theory; unitarity; indirect constraints.

1 Introduction

I am asked to discuss 'Field Theory and the EW Standard Model' in these four lectures. The title encompasses developments of the last 60-70 years. These lectures are happening on the backdrop of the discovery of the Higgs at the LHC [1], the concluding finale of the establishment of the correctness of the Standard Model as the theoretical description of EW interactions. To cover this entire journey in four lectures, clearly I have had to pick and choose a few topics. I have done after sharing a questionnaire with all of you.

I would like to focus on the salient and non negotiable aspects of EW phenomenology which helped establish the $SU(2)_L \times U(1)_Y$ gauge field theory as the correct theory of the EW interactions. In this I will like to tell the story of how requirements of consistency of EW theory itself have guided us in the development of Standard Model (SM), as we know it today, by setting up the goal posts for theory and experiments. I will begin by discussing some aspects of the pre-gauge theory description of weak interactions in terms of a current-current Lagrangian. As we understand today this is the effective theory which results from the $SU(2)_L \times U(1)_Y$ description, when the heavy gauge boson fields have been integrated out. It is interesting to understand the role that various features of this effective description have played in helping us 'infer' the more fundamental theory which is the SM. I will try to point out some of these. I will then begin a discussion of SM as a gauge theory, by first setting up the notation of the SM Lagrangian followed by a somewhat brief discussion of the Higgs mechanism. Then I give a very brief summary of the successes of the SM all the way from its formulation till date. I will then discuss relationship between the particle spectrum of the SM and the twin issues of anomaly cancellation and custodial symmetry. I will then sketch how one can understand the development of the SM as a theory in terms of taming bad high energy behavior of scattering amplitudes. Then will come a discussion of the GIM mechanism and 'prediction' of the mass of the charm quark M_c from the measured mass difference between K_0 and $\bar{K}_0$. This will be followed by a discussion of the experimental measurements which established the EW part of the SM as a quantum gauge field theory based on the gauge group $SU(2)_L \times U(1)_Y$, albeit where the symmetry is broken spontaneously. I will assume essentially that

https://doi.org/10.23730/CYRSP-2017-002.1

people are aware of some of the details of the Spontaneous Symmetry Breaking (SSB) and hence will only sketch it here. As we know establishing the $SU)2)_L \times U(1)_Y$ theory with SSB as the correct theory of EW interactions was done by testing the precision measurements of various EW observables against the predictions for the same including radiative corrections. Inclusion of these radiative corrections is possible only in a renormalisable quantum field theory. In particular I will discuss the history of determination of M_t and M_h from 'indirect' effects on observables through loop corrections. In the last lecture I will discuss various theoretical bounds on the Higgs mass and also the theoretical implications of the observed mass of the Higgs at the LHC [2, 3] for the SM.

2 Preliminaries

2.1 Periodic table of particle physics

The SM stands on the joint pillars of relativistically invariant quantum field theories and gauge symmetries. The SM is a quantum gauge field theory based on the gauge group $SU(3)_C \times SU(2)_L \times U(1)_Y$ which describes the strong and electro-weak(electromagnetic and weak) interactions. The subject matter of these lectures is going to cover only the EW part of the SM. Gauge theory of strong interactions, QCD, will be discussed in a different set of lectures at this school.

As things stand today, the periodic table of the SM is complete. One part of this periodic table are the spin-$\frac{1}{2}$ matter particles: the quarks and the leptons and their anti-particles. Table 1 summarises the details of the currently available information on all the matter fermions.

Table 1: Elementary fermions of the Standard Model, all of spin $\frac{1}{2}$. The three quark colours are indicated explicitly, while leptons are colourless. Electric charges in units of the positron charge, are displayed on the left side. The anti-particles form a similar table with opposite charges.

Quarks	Leptons
$2/3 \quad \begin{pmatrix} u \\ d \end{pmatrix} \quad \begin{pmatrix} c \\ s \end{pmatrix} \quad \begin{pmatrix} t \\ b \end{pmatrix}$	$0 \quad \begin{pmatrix} \nu_e \\ e \end{pmatrix} \quad \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix} \quad \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}$
$2/3 \quad \begin{pmatrix} u \\ d \end{pmatrix} \quad \begin{pmatrix} c \\ s \end{pmatrix} \quad \begin{pmatrix} t \\ b \end{pmatrix}$	
$2/3 \quad \begin{pmatrix} u \\ d \end{pmatrix} \quad \begin{pmatrix} c \\ s \end{pmatrix} \quad \begin{pmatrix} t \\ b \end{pmatrix}$	
$M_u = 2$ MeV	$M_{\nu_1} = 0 - 0.13 \times 10^{-6}$ MeV
$M_d = 5$ MeV	$M_e = 0.511$ MeV
$M_c = 1,300$ MeV	$M_{\nu_2} = 0.009 - 0.13 \times 10^{-6}$ MeV
$M_s = 100$ MeV	$M_\mu = 106$ MeV
$M_t = 173.000$ MeV	$M_{\nu_3} = 0.04 - 0.14 \times 10^{-6}$ MeV
$M_b = 4.200$ MeV	$M_\tau = 1.777$ MeV

Of course, a gauge field theoretic description of the interactions among these elementary particles needs in the SM particle spectrum, also the gauge bosons which would be the carrier of the various interactions. This leads to the second set of members of the 'periodic table' of particle physics, viz. the spin-1 gauge bosons: the photon, W and Z bosons and gluons. Their details are indicated in Table 2.

As we will discuss in detail later, gauge invariance, which guarantees the renormalisability of this theory, would require that all of the gauge bosons should be massless. Not only that, the same invariance

Table 2: Elementary bosons of the Standard Model. There are no separate anti-particles: W^- is the anti-particle of W^+ and the rest are neutral. Q indicates the electromagnetic charge of the boson in units of positron charge.

Electromagnetic and weak (Spin 1)	Strong (Spin 1)	Higgs (Spin 0)
γ (photon) $W^\pm$, Z (weak bosons)	g (gluons)	h (Higgs)
$M_\gamma = 0,\ Q_\gamma = 0$ $M_W = 80.404$ GeV, $Q_W = \pm 1$ $M_Z = 90.1876$ GeV, $Q_Z = 0$	$M_g = 0,\ Q_g = 0$	$M_h = 125.4\pm$ GeV, $Q_h = 0$

would require the matter fermions also to be massless. However, other than the gluon and the photon all the other members of this periodic table (cf. Tables 1 and 2) are patently massive. In fact, it is the mechanism of Spontaneous Symmetry Breaking (SSB), which allows these particles to have non zero masses and helps keep the theory still consistent with gauge invariance. SSB of the EW gauge symmetry via the Higgs mechanism (or Brout-Englert-Higgs mechanism for the purists) [4], is the key ingredient of renormalisable gauge theories of the EW interaction. This requires existence of yet another member of the periodic table, which is the Higgs boson. This too has been included in the list of the SM bosons in Table 2, now that its existence has been established firmly and the discovery awarded a Nobel prize!

2.2 Weak interactions: pre-gauge theory

Fermi's theory of β decay [5], was the blueprint of the early theoretical description of the weak interactions which are responsible not just for the radioactive β decays of nuclei but also for the strangeness conserving and strangeness changing weak decays of the mesons and baryons. This culminated in the famous V-A theory of weak interactions [6, 7]. According to this theory, the μ decay $\mu^- \to \nu_\mu e^- \bar{\nu}_e$ for example, could be described by an effective Hamiltonian

$$\mathcal{H}_{eff}^{\mu\ decay} = -\frac{G_\mu}{\sqrt{2}}\left[J_{\nu e}^{\rho+\dagger} J_{\nu\mu,\rho}^{+} + h.c.\right], \tag{1}$$

where

$$J_{12}^{\rho+} = \bar{\psi}_1 \gamma^\rho (1-\gamma_5)\psi_2 \equiv J_{12}^{\rho CC}. \tag{2}$$

In the same way, the β decay of the neutron could be described by an effective interaction given by

$$\mathcal{H}_{eff}^{\beta decay} = -\frac{G_F}{\sqrt{2}}\left[J_{e\nu}^{\mu+\dagger} J_{\mu,pn}^{+} + h.c.\right], \tag{3}$$

with

$$J_{pn}^{\mu+} = \bar{\psi}_p (1 - 1.26\gamma_5)\gamma^\mu \psi_n \tag{4}$$

In fact, it was established that when written in terms of the quarks which make up the mesons and baryons, all the weak processes could be described in terms of a four fermion, current-current interaction depicted in left panel of Figure 1 which shows a transition $1 \to \bar{2}{+}3{+}4$. For example, the basic transition describing the n decay $n(udd) \to p(uud) + e^- + \bar{\nu}_e$, is given by the current-current interaction depicted

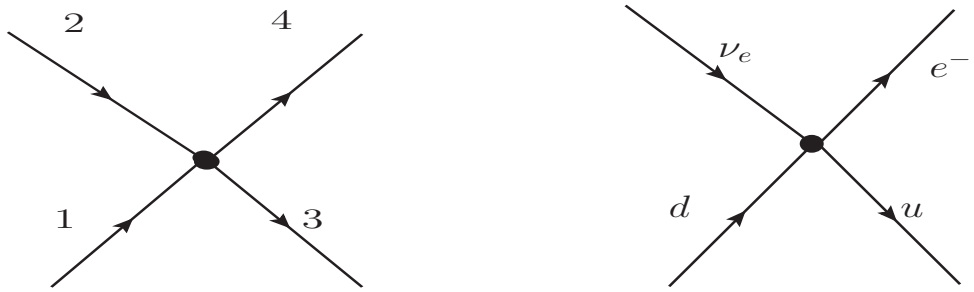


Fig. 1: Generic four fermion interaction responsible for the weak processes (left panel) and the basic process describing the β decay (right panel)

in the right panel. The crux of $V-A$ theory is that only the left chiral fermions are involved in this weak interaction Hamiltonian. The effective Hamiltonian is then written as

$$\mathcal{H}_{eff}^{4fermion} = -\frac{G_\mu}{\sqrt{2}}\left[J_{24}^{\mu+\dagger}J_{31,\mu}^{+} + h.c.\right] = -4\frac{G_\mu}{\sqrt{2}}\left[\left(\bar{\psi}_{3L}\gamma^\mu\psi_{1L}\right)\left(\bar{\psi}_{4L}\gamma_\mu\psi_{2L}\right) + h.c.\right] \tag{5}$$

The appearance of $\psi_L = 1/2(1-\gamma_5)\psi$ in Eq. 5, indicates that only left chiral fermions are involved in this charged weak current. As we will see later, it is this fact that decides the representation of the $SU(2)_L$ gauge group to which the various fermion fields belong.

We understand the electromagnetic interaction in terms of the electromagnetic current $J_\mu^{em} = \bar{\psi}_L\gamma_\mu\psi_L + \bar{\psi}_R\gamma_\mu\psi_R$ and the electromagnetic field A_μ. The corresponding vertex is depicted in the left panel of Fig. 2. Eq. 5 means that one can similarly think of the weak current J_μ^+ (for example) coupled to a charged gauge boson (a weak boson W) W_μ^+. The basic transition brought about by the charged current could then be depicted as shown in the right panel of Fig. 2. The electromagnetic charge of f'

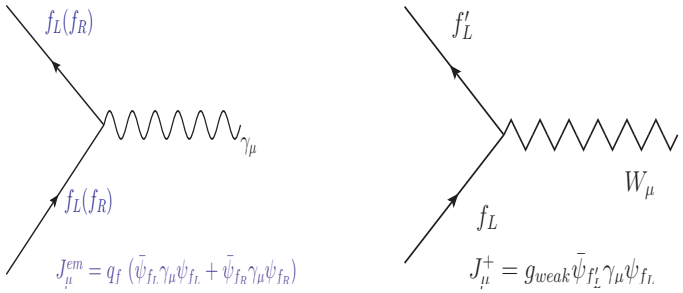


Fig. 2: The left panel shows the usual QED vertex depicting the interaction with the QED gauge boson γ_μ and the right panel shows the generic vertex describing the universal weak interaction among quarks and leptons.

differs from that of f by one unit and in case f is strange quark, the strangeness changes by one unit as well. In that case this current indicates a transition which brings about $\Delta S = \Delta Q = 1$, where S and Q stand for the strangeness and the electromagnetic charge respectively. While the decay of a neutron n involves the current $J_{ud}^{\mu+}$, the decay of Λ for example, involves the current $J_{us}^{\mu+}$. The strength of the four-fermion interaction is then decided by g_{weak} of Fig. 2. Experimentally measured values of G_μ and G_F of Eqs. 1, 3 were somewhat different from each other, though very close, $G_F \sim 0.98G_\mu$. For the effective Hamiltonian for Λ decay, for example, the corresponding coefficient was yet again different from both G_μ, G_F, G_Λ being $0.20G_\mu$ [1]. It was Cabibbo's observation [8] that all this could be consistent with a

[1]The very near equality between G_μ and G_F was an indication that the vector current was not affected by the strong

completely universal charged weak current i.e., a current which has the same strength for the leptons as well as the quarks and also for $\Delta S = 0$ and $\Delta S = 1$ alike, if in case of quarks, the basic charged current in Fig. 2 describes a transition with $f' = u$, $f = d' = d\cos\theta_c + s\sin\theta_c$, with $\sin\theta_c \sim 12^\circ$. This means that the interaction eigenstate d' is a linear combination of the mass eigenstates d and u. Clearly, the orthogonal combination $s' = -d\sin\theta_c + d\cos\theta_c$, is an interaction eigenstate coupling with a $W^\pm$ and a *new* quark with charge $+\frac{2}{3}$. This thus indicates existence of the fourth quark : the charm quark c. As we will see later its existence ensures flavour conservation of the weak neutral currents at tree level automatically. This then helps one understand the experimentally observed suppression of the Flavour Changing Neutral Currents (FCNC) which will be discussed in detail later. Thus the states to be identified with the interaction eigenstates would be:

$$\begin{pmatrix} u' \\ d' \end{pmatrix} = \begin{pmatrix} u \\ d\cos\theta_c + s\sin\theta_c \end{pmatrix}; \quad \begin{pmatrix} c' \\ s' \end{pmatrix} = \begin{pmatrix} c \\ -d\sin\theta_c + s\cos\theta_c \end{pmatrix}$$

At this point let us also mention one more feature of the phenomenology of quark mixing which will be relevant later. In fact, the physics of the $K_0, \bar{K}_0$ mesons not only revealed the existence of suppressed nature of the FCNC but also CP violation in K_0–$\bar{K}_0$ system. This CP violation can also be understood as coming from the above quark-mixing but ONLY if the mixing matrix involves a phase. For this to be possible we have to have at least three generations of quarks. This was noted by Kobayashi-Maskawa [9]. This makes it possible to understand the CP violation observed in the neutral meson system, in the context of a gauge theory of EW interactions, in terms of the mixing in the quark sector. However, this requires existence of at least three generations. Thus one sees that in some sense, the need to understand the observed phenomenology of FCNC and CP violation, in the framework of a gauge theory, predicted the existence of the c and the t quark respectively.

For future reference note that the connection between the mass eigenstates u, d, c, s, t and b and the interaction eigenstates u', d', c', s', t' and b' is given by $u' = u, c' = c, t' = t$ and

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}, \tag{6}$$

where V_{ud} etc. are elements of the CKM matrix $\mathbf{V}$ Refs. (cf. [8]– [9]). This describes the interaction eigenstates in terms of the mass eigenstates.

At this point let us also note that the same four fermion interaction that describes the decay $\mu^- \to e^- + \bar{\nu}_e + \nu_\mu$ can also describe, for example, the scattering processes such as $\nu_\mu + e^- \to \nu_e + \mu^-$, corresponding to $1 = e^-, 2 = \nu_\mu, 3 = \nu_e$ and $4 = \mu^-$ in the left panel of Fig. 1. The same effective Hamiltonian as in Eq. 5 then also describes this scattering process as well. If one calculates the total cross-section one gets,

$$\sigma_{\rm tot} = \frac{G_\mu^2 s}{\pi} = \frac{2G_\mu^2 m_e E_{\nu_\mu}}{\pi}. \tag{7}$$

This linear rise of scattering cross-section with s, the square of the centre of mass energy or alternatively E_{ν_μ}, is a reflection of the 'pointlike' nature of the Fermi interaction of Eq. 5. It can be seen, by doing a partial wave analysis of the scattering amplitude, that this behaviour implies violation of unitarity when $\sqrt{s} \geq 300$ GeV. Of course, in practical terms it corresponds to a $E_{\nu_\mu} \geq 10^8$ GeV and hence perhaps not very relevant. However, it is the principle that matters. A cure to this problem of the current -current interaction was indeed offered by postulating the existence of a massive, charged boson (called the weak-boson $W^\pm$) by Schwinger. This is the same $W^\pm$ we have already introduced while

interactions of the n and p and is the same for e^- to ν transition as for n to p. This was called the 'Conserved Vector Current hypothesis' (CVC). In all the discussions regarding the mixing angle, we are referring to the coefficient of this conserved vector part of the current at the hadron level.

writing the weak vertex in Fig. 2. Thus the point interaction of Eq. 5 can be understood as an interaction resulting from the exchange of a $W^{\pm}$ boson, in the limit of the said mass M_W being much bigger than all the energies in the system. This is depicted in Fig. 3. The observed short range of the weak

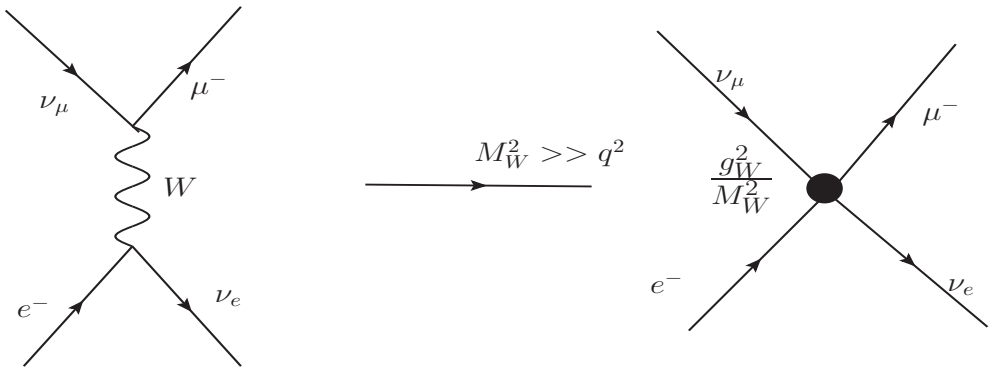


Fig. 3: Contact interaction resulting from $M_W \to \infty$ limit

force causing the β decay, indicated that the $W^{\pm}$ boson is massive, unlike the photon mediating the electromagnetic interaction which is massless. The success of the effective Hamiltonian of Eq. 5 implies a lower bound much bigger than MeV and hence $\sim \mathcal{O}$ (GeV). To summarize, we see that the requirement that unitarity bound be respected, indicates the existence of a massive charged vector boson $W^{\pm}$ and the four-fermion weak interactions can be understood as caused by an exchange of this massive boson. The 'massive' nature of the exchanged boson was also consistent with the observed 'short' range of the weak interactions. However, if it is a gauge boson, then the massive nature will also break gauge invariance! Further, the massive nature of the gauge boson causes problems such as bad high energy behavior of scattering amplitudes as well as non renormalisability of the theory. How a massive gauge boson is to be accommodated in the framework of a gauge theory is going to be the topic of discussion in the next section.

2.3 Observations meet predictions of the SM

Before beginning with a discussion of details of a gauge theory, let us just briefly take a look how the establishment of the SM has been a synergistic activity between theoretical and experimental developments. We saw already how the form of the pre-gauge theory, effective Hamiltonian description of weak interactions, obtained phenomenologically from the data hinted at a possible gauge theoretic description of the same. Equally interesting are the hints at existence of new particles given by the theory. While some of the members of this periodic table, like the μ, were unlooked for and some like the ν were met with quite a bit of disbelief when postulated theoretically, for most of the recent additions their existence and in some cases even their masses were predicted if the EW interactions were to be described by a renormalisable gauge theory.

In fact, the existence of strange particles which contain the strange quarks, coupled with experimental features such as the suppression of the FCNC in EW processes alluded to before, indicated the existence of the charm quark, as already indicated above. Further, the small mass difference between K_L and K_S (or alternatively the K_0–$\bar{K}_0$ mixing) could be used to obtain an estimate of its mass. Accidental discovery of some members of the third lepton and quark family, combined with the requirement of anomaly cancellation, an essential feature for a renormalisable theory, meant that the remaining members of the same family had to exist. Hence t and the ν_τ were hunted for very actively once the b and the τ made their appearance! The properties of a renormalisable quantum field theory were the essential reasons behind the belief in these predictions. The mass of the t quark could also be predicted in the SM,

using experimental information on neutral B meson mixing and properties of the Z boson, as we will see below.

The story is not very different for the EW gauge bosons. As was already mentioned, requiring consistency of the pre gauge theory description of the weak interactions with unitarity, had indicated a nonzero mass for the charged $W^{\pm}$ but had not indicated what the mass would be, except that it should be much larger than the typical energy scales involved in the weak decays $\sim$ MeV. It is the unified description of the EW interactions of the Glashow-Weinberg-Salam (GSW) model [11] that actually gave a lower limit on its mass. Note that the correctness of the $V-A$ nature of weak interactions and pure vector nature of the electromagnetic interactions predicted existence of a neutral boson other than the photon γ. In the GSW model, the masses of the W and the new Z boson required in the unified EW theory, were all predicted in terms of the life time of the μ and the weak mixing angle θ_W which was a free parameter in the model. This could be determined from measurements of rates of various weak processes.

Not just this, the SM also predicted existence of yet another boson, this time spin 0; viz. the Higgs boson. The mass of the said Higgs boson, however, is a free parameter in the framework of the SM. Comparisons of the EW observables with precision measurements can constrain the Higgs mass through the corrections caused by the loop effects which can be computed in a renormalisable quantum field theory. One can also put limits on this parameter from theoretical considerations of consistency of the SM as a field theory at high scales: the triviality and vacuum stability, all to be discussed in the lectures.

Let us discuss in detail the case of the t quark which is quite interesting. The existence of the t quark and the information on its mass came from a variety of theoretical and phenomenological observations in flavour physics and physics of the W/Z bosons. As already mentioned the explanation of the experimentaly observed CP violation in terms of the quark mixing matrix requires at least three generations of quarks. This mixing is described by the famous CKM mixing matrix Refs. [8]– [9]. So in that sense existence of the t and b was indicated by this observation.[2] Experimental manifestation of B_0–$\bar{B}_0$ oscillations at the ARGUS experiment [10] was a harbinger of the presence of the t quark. Further indications for the expected mass actually came from precision measurements of many EW observables, ie. properties of the Z and the W boson and the quantum corrections caused to them by loops containing top quarks.

Experimental observation of the t quark at the Tevatron [12], with a mass value consistent with the implications of the EW precision measurements, provided a test of the description, at loop level, of EW interaction in terms of an $SU(2)_L \times U(1)_Y$ gauge field theory with SSB. Fig. 4 shows, by open circles, evolution with time of the values of the top mass extracted *indirectly* by comparing the measured EW parameters with the SM predictions. Also shown are the 95% c.l. upper limits from direct searches from the e^+e^- experiments (solid line) and from $p\bar{p}$ experiments (the dashed line). In the last part of the plot the solid triangles show the mass of the top quark measured directly at the Tevatron and the 'indirectly' extracted values of M_t at the same time. The remarkable agreement between directly measured and the 'indirectly' extracted values around the time of the discovery, was a test of the SM at loop level.

Once this was achieved, the same information could be used to obtain constraints on the Higgs mass, now looking at quantum corrections to the W, Z mass as well as to the Z couplings, caused by loops containing the Higgs boson. Finally finding a Higgs boson in 2012 [2] with a mass consistent with these constraints was the biggest success of the SM [3] Fig. 5 reproduced from the Gfitter webpage [14] illustrates this. The various dark and light shaded regions correspond to 68% and 95% c.l. contours in all cases. The green bands between the vertical and horizontal lines indicate experimentally measured values of M_t and M_W. The region shaded in blue (the long and narrow ellipses) indicates the region allowed in

[2] The requirement of anomaly cancellation for the gauge theory of EW interactions to be renormalisable, further indicated existence of an additional generation of leptons, τ, ν_τ as well.

[3] Knowledge of QCD, the part of the SM which we are not discussing in these lectures, was essential in making precision predictions for the Higgs signal and hence to this mass determination!

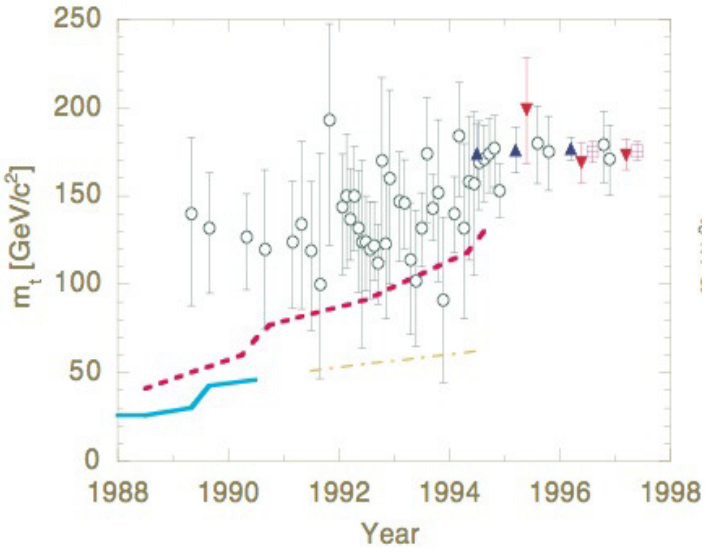


Fig. 4: Comparison of the limits on the mass of the top quark from direct searches at the e^+e^- collider (solid line) and the hadronic colliders (red dashed line) with the indirect limits, indicated by open circles, coming from precision EW measurements as a function of time. The dot dashed line is an indirect lower limit obtained from the observed rates of inclusive W/Z production in $p\bar{p}$ colliders.The solid triangles indicate directly measured values of the mass of the observed t-quark. This is taken from [13].

the M_t–M_W plane, by fits of the SM prediction for precision measurements of EW observables where the Higgs mass [2] information is used. The big elliptical regions, one of them open at one end, shaded in light and dark grey, are the ones allowed when none of the mass measurements are used as input and one lets the EW precision data choose the best fit values. Consistency of the values obtained in these fits with each other and with the experimental measurements indicated by the small oval with dark and pale green regions, leaves us with no doubt about the correctness of the SM. This tests the correctness of quantum corrections to M_W coming from the loops containing the t and h; hence of the quantum field theoretic description of the EW interactions as a gauge theory.

Alongside this spectacular testimonial of the correctness of the EW part of the SM, is also the equally impressive demonstration of a highly accurate description of all the CP violating phenomena in terms of the flavour mixing in the quark sector. In the three flavour picture the 3×3 CKM matrix is unitary. Making detailed fits of theoretical predictions to a large variety of data on meson mixing and decays, to determine the elements of the CKM matrix with high precision, is an involved exercise as it requires a synthesis of a variety of theoretical tools and high precision data. These elements are parameterised in terms of two parameters : $\bar{\rho} - \bar{\eta}$ [3]. Fig. 6 taken from PDG-2015 shows the constraints in the $\bar{\rho} - \bar{\eta}$ plane from a variety of measurements around the global fit point. Various shaded areas indicate the regions allowed at 95% c.l. from a given measurement. The unitarity of the CKM matrix is indicated by the fact that the 'tip' of the unitarity triangle lies in the small intersection region allowed by all the various measurements. Since for many of these observables their relationship with the parameters of the SM is given by loop computations, this success too provides a test of the SM as a quantum gauge field theory.

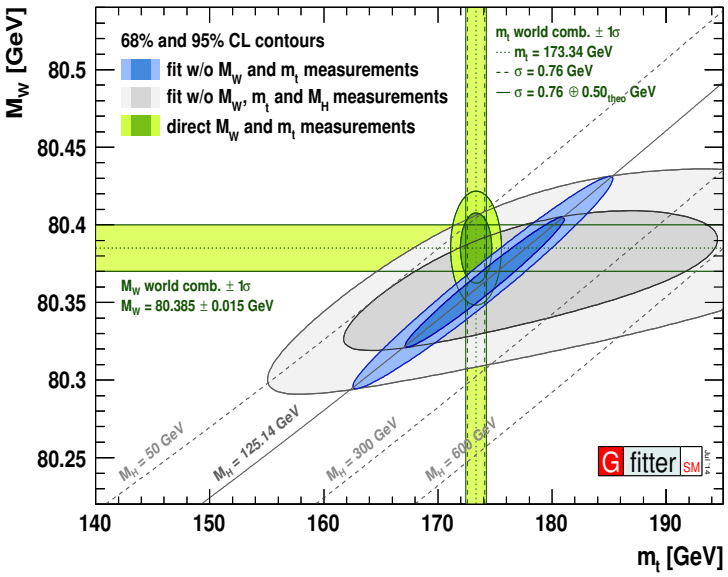


Fig. 5: Testing of the SM at loop level from mass measurements. $M_W - M_t$ values consistent with the EW precision data with and without using measured value of M_t, M_W and M_h as inputs. Taken from [14]

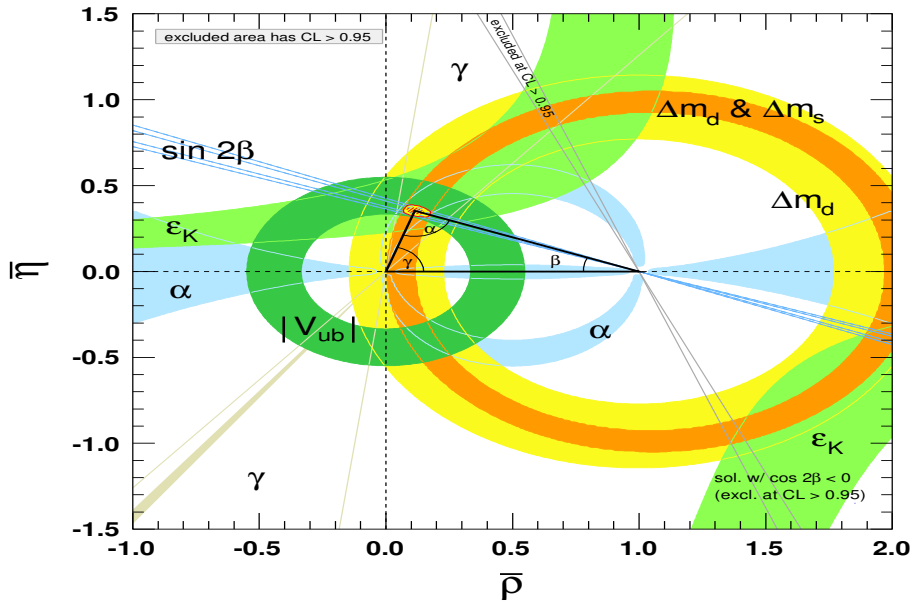


Fig. 6: Constraints in the $\bar{\rho} - \bar{\eta}$ plane from a variety of measurements around the global fit point. Taken from [3].

3 $SU(2)_L \times U(1)_Y$ gauge theory

3.1 Gauge principle

Gauge principle is the basis of the theoretical description of three of the fundamental interactions *viz.* strong, weak and electromagnetic, among the quarks, leptons and the force carrying gauge bosons. QED is the first gauge theory to be established. We therefore can begin our discussion of gauge theories, by looking at QED: a theory of a Dirac fermion field $\psi(x \equiv \vec{x}, t)$ of charge e. For a free Dirac fermion of mass m the Lagrangian density consists of the kinetic term supplemented with the mass term and is

given by

$$\mathcal{L}_f = i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi.$$

However, this Lagrangian density is not invariant under a local $U(1)$ gauge transformation,

$$\psi(x) \to e^{i\alpha}\psi(x), with\ \ \alpha = \alpha(x). \tag{8}$$

Note that this non invariance of the Lagrangian density is true only for the local gauge transformation with $\alpha = \alpha(x)$. To construct a gauge invariant Lagrangian density, one needs to introduce a vector field A_μ and generalise the derivative $\partial_\mu \to \partial_\mu + iq_f|e|A_\mu$ where q_f is the charge of the fermion in units of positron charge $|e|$. Thus for the electron, the covariant derivative is

$$D_\mu = \partial_\mu - ieA_\mu \tag{9}$$

Combining this generalization of the kinetic term for the fermion, with the gauge transformation of the vector field

$$A_\mu \to A_\mu + \frac{1}{e}\partial_\mu\alpha(x), \tag{10}$$

one can show that $(\partial_\mu - ieA_\mu)\psi \to e^{i\alpha(x)}(\partial_\mu - ieA_\mu)\psi$. Thus, under the gauge (phase) transformation of the fermion field, the vector field too has to transform with the same transformation parameter $\alpha(x)$. Note now that the Lagrangian density

$$\begin{aligned}
\mathcal{L}_{QED} &= i\bar{\psi}\gamma^\mu D_\mu\psi - m\bar{\psi}\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} = i\bar{\psi}\gamma^\mu(\partial_\mu - ieA_\mu)\psi - m\bar{\psi}\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \\
&= i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + e\,\bar{\psi}\gamma^\mu\psi A_\mu \\
&= \mathcal{L}_f + \mathcal{L}_{gauge} + e\,\bar{\psi}\gamma_\mu\psi A^\mu = \mathcal{L}_f + \mathcal{L}_{gauge} + \mathcal{L}_{int},
\end{aligned} \tag{11}$$

with $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, is gauge invariant. Note that a mass term for the vector field *viz.* $M_A^2 A_\mu A^\mu$ will break this gauge invariance (cf. Eq. 10). Further, this Lagrangian density is just the sum of three Lagrangian densities: $\mathcal{L}_f$ for the free fermion field ψ of mass m as given by the first two terms in the third line of Eq. 11, $\mathcal{L}_{gauge}$ for free *massless* gauge field A_μ given by the third term and the interaction term $\mathcal{L}_{int}$ being given by the last one. Note that the form of the interaction of the fermion with the gauge field is *completely fixed* by the form of the covariant derivative D_μ. Further, the mass term

$$m\bar{\psi}\psi = m(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L)$$

will not be invariant under an $U(1)$ local gauge transformation similar to that given by Eq. 8 if, for example, the left and right chiral fermions have different $U(1)$ charges. This will be the case with $U(1)_Y$ gauge group of the Standard Model, as we will see very soon.

Note also the interaction term given by:

$$\mathcal{L}_{int}^{QED} = e\bar{\psi}\gamma^\mu\psi A_\mu = eJ^{\mu,em}A_\mu. \tag{12}$$

The current J_μ^{em} of Eq. 12 is the vector bilinear constructed out of the fermion fields ψ and $\bar{\psi}$. As opposed to this, the weak current $J_{12}^{\pm\mu}$ defined in section 2.2 contains a linear combination of both the vector and axial vector bilinears. This phenomenologically ascertained form of the weak current therefore pointed already towards a gauge theory of weak interactions albeit with parity violation. The form of this chirality conserving current indicated existence of two charged vector bosons which however couple only to left chiral fermions. Thus the V–A form of the current-current interaction already gives indications about the representation of this gauge group, to which different types of fermions should belong since, as seen above, in a gauge theory it is this representation that decides the interaction of the fermions with the vector gauge bosons. The similarity and the differences in the nature of the weak and electromagnetic

current and description of electromagnetic interactions in terms of a $U(1)$ gauge theory, paved the way towards an unified description of electromagnetic and weak interactions as the electro-weak gauge theory based on the gauge group $SU(2)_L \times U(1)_Y$.

Before we formally write down the complete Lagrangian density for the EW part of the SM, let us discuss the generalisation of the above discussion to non ablian gauge transformations. To that end let us begin by summarising some of the relevant observations for QED, which we have stated above. The local phase transformations given by Eq. 8 form an unitary group and is called $U(1)$. The Lagrangian density of matter fields is invariant under this $U(1)$ transformation only if there exists a vector field which simultaneously transforms with the same transformation parameter and the matter field interacts with this vector field in a specific manner. We consider now, a generalisation of this simple symmetry transformation of Eq. 8 to a case where matrix valued analogues of this simple phase transformations act on a set of fields and again the elements of the matrices can depend on the space time coordinates of the point : $\vec{x}, t$. Again, invariance of the matter Lagrangian density under this local transformation requires a set of spin 1, vector fields which transform under the local gauge transformation according to a generalisation of Eq. 10 in addition to a modification of the kinetic term of the matter fields by replacing ∂_μ by the covariant derivative D_μ as done above. Thus there exists now a multiplet of gauge bosons. Another curious property of the Lagrangian density involving these gauge fields is that even in absence of matter fields and interactions, the equations of motion are non linear. This in turn means that the associated spin 1 particles interact with each other in the absence of matter. Further, unlike the phase transformations of the QED, these matrix valued transformations do not commute with each other. Hence these generalized gauge theories are also called non-abelian gauge theories.

Lagrangian density of a free, massless non-Abelian gauge theory is given by

$$\mathcal{L}_{nonabelian} = -\frac{1}{4} F^a_{\mu\nu} F^{a,\mu\nu} \tag{13}$$

with

$$F^a_{\mu\nu} = \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g f^{abc} W^c_\mu W^c_\nu \tag{14}$$

Here f^{abc} are structure constants which are specific to each gauge group defined by,

$$[T^a, T^b] = i f^{abc} T^c, \tag{15}$$

T^a being the generators of the gauge transformation. f^{abc} are called the structure constants. T^a are called generators because, in general if Φ represents a matter field (spin $\frac{1}{2}$ or spin 0) transforming according to a representation T_{IJ} of the gauge group then

$$\Phi_I \rightarrow exp^{-ig(T^a)_{IJ}\alpha^a(x)} \ \Phi_J, \tag{16}$$

where g is the coupling constant. The repetition of index a indicates sum over all the generators of the transformation. The covariant derivative is then given by

$$D_\mu \Phi_I = \partial_\mu \Phi_I - i g V^a_\mu (T^a)_{IJ} \Phi_J, \tag{17}$$

where V^a_μ denote the associated spin 1 vector fields. The kinetic term for the matter fields, defined in terms of the D_μ along with the one for massless gauge fields given by Eq. 13, are both invariant under the gauge transformation if the gauge field also transforms as

$$V^a_\mu \rightarrow V^a_\mu + \frac{1}{g}\partial_\mu \alpha^a + f^{abc} V^a_\mu \alpha^c. \tag{18}$$

Again the couplings of the matter particles with the gauge bosons V^a_μ, are then given by the kinetic term written down using the covariant derivative given by Eq. 17, just like we did in Eqs. 11 and 12. We can

then write down currents J^V_μ analogous to J^{em}_μ of Eq. 12. This is completely determined once we specify the gauge group, i.e., T^a, the representation of the gauge group to which the matter particles belong and the coupling g.

When $a = 1$, i.e., when there exists only one gauge boson, these gauge transformations and covariant derivative given by Equations 16–18 reduce to those for simple phase transformation corresponding to the $U(1)$ case, viz., Equations 8–10. For the case where a is different from 1, because of the commutator relation, the normalisation of the charge g is fixed for all the representations. For $U(1)$ gauge transformation on the other hand the normalisation of the charge can be different for different representations. For future reference, let us also note here that for the $SU(2)$ gauge group we have

$$T^a = \frac{\tau^a}{2} \ and \ f^{abc} = \epsilon^{abc}, \quad a = 1 - 3$$

where $\tau^a, a = 1 - 3$ are the Pauli matrices and ϵ^{abc} is the constant, completely antisymmetric tensor. Hence, for $SU(2)$ the index a takes values 1–3 in Eq. 16.

3.2 GSW model

Let us first write down the gauge boson and matter particle content for the GSW model along the interactions among all these. The gauge group for the GSW model is $SU(2)_L \times U(1)_Y$. The subscript L means that the gauge transformations corresponding to this gauge group are non trivial ONLY for the left chiral(handed)[4] fermions and the right chiral fermions remain unchanged under it. The direct product means that these two groups are independent, i.e., the left handed fermions belonging to a given representation of $SU(2)_L$ will all have the same value of the charge under $U(1)_Y$. Thus ONLY the left chiral fermions belong to the nontrivial representation of the $SU(2)_L$ group and the right chiral fermions are singlets under the $SU(2)_L$ gauge group. Therefore these have NO interactions with the gauge bosons corresponding to the $SU(2)_L$ gauge group.

3.2.1 Particle content and Currents of the GSW model

For the $SU(2)$ group, each representation is labelled by two quantum numbers T_L and T_{3L}, where T_L takes integral or half integral values: $0, 1/2, 1, 3/2$... etc. and for a given T_L, T_{3L} takes values from $-T_L$ to $+T_L$ in steps of 1. Thus number of fields belonging to representation labelled by T_L is then $2T_L + 1$. For singlet representation $T_L = 0$ and for the doublet it is $1/2$. Thus a doublet of $SU(2)_L$ contains two members with $T_{3L} = \pm 1/2$. The gauge bosons belong to the $T = 1$ representation (called the adjoint representation) and hence they are three in number called $W^a_\mu, a = 1 - 3$. The $U(1)_Y$ gauge group has only one generator like the QED case discussed above. We denote the corresponding single gauge boson B_μ. The corresponding current is J^Y_μ and the charge is called "hypercharge". The electromagnetic charge of a charged fermion is independent of its chirality. On the other hand, the two left chiral fermions of different electromagnetic charges have to have the same $U(1)_Y$ charge. Thus it is clear that the $U(1)_Y$ can not be identified with $U(1)_{em}$, i.e., the hypercharge is different from the electromagnetic charge. Thus $U(1)_{em}$ arises out of a linear combination of $U(1)_Y$ and a $U(1)$ subgroup of$SU(2)_L$.

First let us discuss the physics in terms of $W^a_\mu, a = 1 - 3$ and B_μ. The gauge groups, the corresponding spin-1 huge bosons and the couplings are indicated in Table 3. As we will see in a minute, if the left handed fermions belong to the doublet representation of $SU(2)_L$, the corresponding charge changing gauge current J^W_μ we would construct from the covariant derivative, has the same form as the J^{CC}_μ of Eq. 2, of the $V - A$ current Lagrangian describing the charge changing weak interactions. Let $\frac{Y}{2}$ denote the charge of the fermion under the $U(1)_Y$ gauge group. The corresponding transformation is given by

$$\psi \rightarrow e^{-i(g_1 Y/2)\alpha_Y(x)} \ \psi \tag{19}$$

[4]The word handedness and chirality can be used interchangeably for massless fermions.

Gauge Group	Gauge Boson Fields	Coupling
$SU(2)_L$	W^a_μ, $a = 1, 2, 3$	g_2
$U(1)_Y$	B_μ	g_1

Table 3: Gauge group, gauge bosons and couplings for the GSW model

whereas, for a $SU(2)_L$ doublet the gauge transformation is given by

$$\Psi = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \rightarrow \Psi' = e^{-ig_2(\tau^a/2)\alpha^a(x)}\Psi. \tag{20}$$

f_1 and f_2 are the $T_{3L} = \pm 1/2$ members of this doublet Ψ respectively. $\tau^a/2$ are the generators T^a for the 2-dimensional fundamental representation.

The fermion content of the GSW model can then be written as shown in Table 4. All the left chiral

Quarks	Leptons
$\begin{pmatrix} u \\ d \end{pmatrix}_L$ $\begin{pmatrix} c \\ s \end{pmatrix}_L$ $\begin{pmatrix} t \\ b \end{pmatrix}_L$	$\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L$ $\begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}_L$ $\begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}_L$
u_R, c_R, t_R d_R, s_R,b_R	e_R, μ_R, τ_R
+anti-quarks	+ anti-leptons

Table 4: The fermions and the representation of $SU(2)_L$ to which they belong.

fermions belong to the doublet representation, with the up-type quarks and neutrinos having $T_{3L} = 1/2$ and d-type quarks and negatively charged leptons having $T_{3L} = -1/2$. Note that according to this there are no right handed neutrinos in the particle spectrum of the SM. The colour gauge group $SU(3)_c$ commutes with the electroweak gauge group : $SU(2)_L \times U(1)_Y$. Hence the electroweak interactions of a quark are independent of its colour. Therefore we suppress here the colour index.

As already discussed $U(1)_{em}$ is a linear combination of $U(1)_Y$ and a $U(1)$ subgroup of $SU(2)$. This is really the essence of Electro-Weak unification and is embodied in Glashow's observation:

$$Q_f = T_{3L} + Y/2. \tag{21}$$

Here Q_f is the electromagnetic charge in units of $|e|$, where e is electron charge, T_{3L} and $Y/2$ denote the $SU(2)_L$ and $U(1)_Y$ charges respectively. Writing the electromagnetic charge as a linear combination of T_{3L} and the hyper-charge Y, embodies the fact that the carrier of electromagnetic interactions, the photon A_μ will appear as a linear combination of the neutral vector boson W^3_μ and the $U(1)_Y$ gauge

boson B_μ. We can discuss this mixing without making any explicit reference to the Higgs sector. This is what we will do first and then summarise the details of the SSB. Note that the three gauge boson fields $W^1_\mu, W^2_\mu, W^3_\mu$: all couple only to left handed fermions and B_μ couples to both the left handed and right handed fermions. B_μ and W^3_μ mix, giving one zero mass eigenstate γ. One then identifies the other one with a new neutral vector boson called Z. One can schematically represent this as shown in the diagram in Fig. 7. Note here that one can discuss this simply at the level of currents which give interactions among matter and gauge bosons in terms of the gauge principle enunciated in Section 3.1, without making any reference to a specific model which will generate these mixing and masses. The essence of this mixing

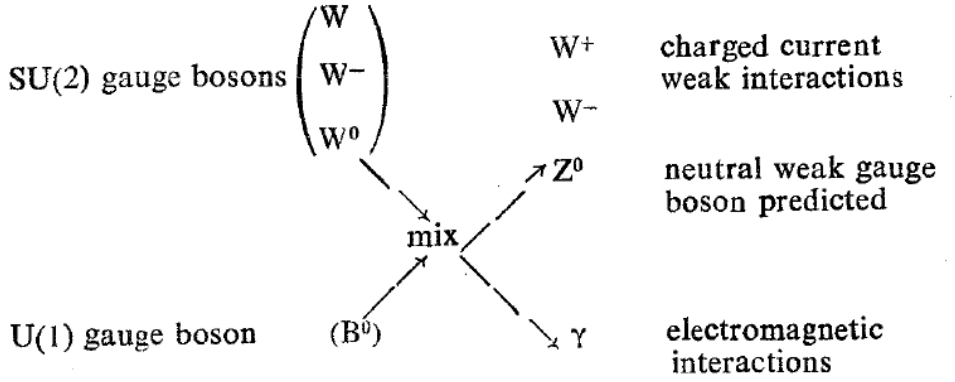


Fig. 7: A schematic description of mixing between the W^3_μ and B_μ. This is taken from [15].

is to define two fields A_μ and Z_μ as a linear combination of B_μ and W^3_μ as:

$$A_\mu = cos\theta_W B_\mu + \sin\theta_W W^3_\mu, \qquad Z_\mu = -\sin\theta_W B_\mu + \cos\theta_W W^3_\mu \tag{22}$$

Here, θ_W, called the 'weak mixing angle', is just an arbitrary parameter denoting the mixing between the W^3_μ and B_μ. To see how the electric charge e is related to g_1, g_2 and $\sin\theta_W$, let us construct the currents J^W_μ and J^Y_μ the way electromagnetic current was constructed in Section 3.1. To do this we need to know the Y values for the different fermion fields written in Table 4. Let us consider a single generation of leptons: e^-, ν_e. Eq. 21 means that the lepton doublet $\mathcal{L}^1_L = \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L$ has $Y = -1$ and $e_{1R} = e_R$ which is an $SU(2)_L$ singlet has to have $Y = -2$. Let us indicate the three lepton doublets written in the last three rows of the Table 4 by $\mathcal{L}^i$ with $i = 1, 3$ respectively. We also use $\mathcal{Q}^i_L$ with $i = 1, 3$ to indicate the doublets $\begin{pmatrix} u^i \\ d^i \end{pmatrix}_L$ where $u^1 = u, d^1 = d$ etc., as written in the first three rows of the same table. For the quark doublets $\mathcal{Q}^i$ the hypercharge Y has value $1/3$. For all the right handed quarks the hypercharge is twice the quark charge and $Y = 2Q_q$, since the value of T_{3L} is zero for all the right handed fields.

Following the discussions in Section 3.1, let us start from the kinetic part of the Lagrangian for all the fermions in Table 4, to construct the physical currents of the GSW model. For the quarks it is simplest when written in the gauge eigenstate basis $u'^i, d'^i, i = 1, 3$. The kinetic term for a fermion field ψ is given by

$$\mathcal{L}_{\text{fermionkin}} = i\bar{\psi}_L \not{\partial}\psi_L + i\bar{\psi}_R \not{\partial}\psi_R, \tag{23}$$

For the $SU(2)_L \times U(1)_Y$ gauge theory the $\not{\partial}$ is to be replaced by the covariant derivative. This can be written in terms of the hyper charges for the fermions given in the earlier paragraph. For a fermion f which is a member of the doublet Ψ this is given by:

$$\partial_\mu \Psi_L \to D_\mu \Psi_L = \partial_\mu \Psi_L - i\frac{g_1 Y_\Psi}{2} B_\mu \Psi_L - i g_2 W^a_\mu \frac{\tau^a}{2} \Psi_L. \tag{24}$$

where $\Psi_L = \mathcal{L}^i_L, \mathcal{Q}_L$ and Y_Ψ is the hypercharge for the doublet Ψ. For the case of $SU(2)_L$ singlets the covariant derivative is given by

$$D_\mu f_R = \partial_\mu f_R - i\frac{g_1 Y_{f_R}}{2} B_\mu f_R. \tag{25}$$

The kinetic terms for all the fermions can be written as:

$$\mathcal{L}_{\rm fermkin} = \sum_{i=1}^{3} \left[i\mathcal{L}^i_L \not{D} \mathcal{L}^i_L + i e^i_R \not{D} e^i_R + i\mathcal{Q}'^i_L \not{D} \mathcal{Q}'^i_L + i u'^i_R \not{D} u'^i_R + i d'^i_R \not{D} d'^i_R \right]. \tag{26}$$

Since there are no right handed neutrinos in the strictest version of the SM, for the lepton sector the mass basis and interaction basis are the same. Using the expressions for the covariant derivative D_μ of Eqs. 24, 25, along with Eq. 6, we find the interaction Lagrangian to be

$$\Delta\mathcal{L}_{\rm int} = \frac{1}{2} g_1 J^{\mu\, Y} B_\mu + g_2 \Big(\frac{1}{2\sqrt{2}} (J^{\mu +} W^+_\mu + J^{\mu -} W^-_\mu) + J^{\mu\, 3} W^3_\mu \Big) \tag{27}$$

where:

$$\begin{aligned}
J^{\mu +} &= 2\big(\bar{\nu}^i_L \gamma^\mu e^i_L + \bar{u}^i_L \gamma^\mu \mathbf{V}_{ij} d^j_L\big) \ , J^{\mu -} = (J^{\mu +})^\dagger, \\
J^{\mu Y} &= -\bar{\nu}^i_L \gamma^\mu \nu^i_L - \bar{e}^i_L \gamma^\mu e^i_L - 2\bar{e}^i_R \gamma^\mu e^i_R + \frac{1}{3}\bar{u}^i_L \gamma^\mu u^i_L + \frac{1}{3}\bar{d}'^i_L \gamma^\mu d'^i_L + \frac{4}{3}\bar{u}^i_R \gamma^\mu u^i_R - \frac{2}{3}\bar{d}'^i_R \gamma^\mu d'^i_R, \\
J^{\mu 3} &= \frac{1}{2}\bar{\nu}^i_L \gamma^\mu \nu^i_L - \frac{1}{2}\bar{e}^i_L \gamma^\mu e^i_L + \frac{1}{2}\bar{u}^i_L \gamma^\mu u^i_L - \frac{1}{2}\bar{d}'^i_L \gamma^\mu d'^i_L, \\
W^\pm_\mu &= \frac{1}{\sqrt{2}}(W^1_\mu \mp i W^2_\mu).
\end{aligned} \tag{28}$$

The couplings must now be rewritten so that one linear combination of B_μ, W^3_μ couples to the electromagnetic current and an orthogonal one couples to $J^{\mu\, 3}$. For this purpose we may ignore the terms in $\Delta\mathcal{L}$ depending on $W^\pm$. For the remaining part, we may think of the physical fields A_μ, Z_μ as the result of a rotation in the B_μ, W^3_μ plane, as already discussed in Eq. 22. We write the inverse rotation:

$$W^3_\mu = \cos\theta_W Z_\mu + \sin\theta_W A_\mu, \qquad B_\mu = -\sin\theta_W Z_\mu + \cos\theta_W A_\mu \tag{29}$$

Inserting into the Lagrangian Eq. 27, we find:

$$\Delta\mathcal{L}(B_\mu, W^3_\mu) = \left[\frac{1}{2} g_1 \cos\theta_W \, J^{\mu\, Y} + g_2 \sin\theta_W \, J^{\mu\, 3} \right] A_\mu + \left[-\frac{1}{2} g_1 \sin\theta_W \, J^{\mu\, Y} + g_2 \cos\theta_W \, J^{\mu\, 3} \right] Z_\mu \tag{30}$$

The expression in the first square bracket in Eq. 30 must be equal to $e J^{\mu\,\rm em} A_\mu$ where e is the unit of electric charge and $J^{\mu\,\rm em}$ is given by an expression for all the charged fermions according to Eq.12 and can be written as

$$J^{\rm em}_\mu = -\bar{e}^i_L \gamma_\mu e^i_L - \bar{e}^i_R \gamma_\mu e^i_R + \frac{2}{3}\Big(\bar{u}^i_L \gamma_\mu u^i_L + \bar{u}^i_R \gamma_\mu u^i_R\Big) - \frac{1}{3}\Big(\bar{d}'^i_L \gamma_\mu d'^i_L + \bar{d}'^i_R \gamma_\mu d'^i_R\Big). \tag{31}$$

This can happen *only if*

$$e = g_1 \cos\theta_W = g_2 \sin\theta_W \tag{32}$$

It follows that:

$$\tan\theta_W = \frac{g_1}{g_2}, \qquad e = \frac{g_1 g_2}{\sqrt{g_1^2 + g_2^2}} \tag{33}$$

Inserting this into Eq. 30 we learn that the coupling of the Z-boson is:

$$\frac{1}{\sqrt{g_1^2 + g_2^2}} \left(-\frac{1}{2} g_1^2 J^{\mu\, Y} + g_2^2 J^{\mu\, 3} \right) Z_\mu \equiv g_z J^{\mu\rm NC} Z_\mu \tag{34}$$

Thus the weak neutral current is given by:

$$g_z J_\mu^{\rm NC} = \frac{1}{\sqrt{g_1^2+g_2^2}}\left(-\frac{1}{2}g_1^2 J_\mu^Y + g_2^2 J_\mu^3\right) \tag{35}$$

where g_z is the coupling constant we associate to the Z-boson. This is a convention, because only the combination $g_z J_\mu^{\rm NC}$ appears in formulae. For convenience we choose:

$$g_z = \frac{g_2}{\cos\theta_W} = \sqrt{g_1^2+g_2^2} \tag{36}$$

With this, the weak neutral current is:

$$\begin{aligned} J_\mu^{\rm Z} = J_\mu^{\rm NC} &= -\frac{1}{2}\frac{g_1^2}{g_1^2+g_2^2}J_\mu^Y + \frac{g_2^2}{g_1^2+g_2^2}J_\mu^3 \\ &= -\frac{1}{2}\sin^2\theta_W J_\mu^Y + \cos^2\theta_W J_\mu^3 = J_\mu^3 - \sin^2\theta_W J_\mu^{\rm em} \end{aligned} \tag{37}$$

where we have written two different forms that are both useful.

Taking a look at the first of Eqs. 28 show us that the charged currents $J^{\mu\pm}$ involve only the left chiral fermions and have the so called V(ector)−A(xial vector) structure. $J_\mu^{\rm em}$ given by Eq. 31 has pure vector nature. Eqs. 28 and 37 clearly show that, unlike the $W^\pm$ bosons, the Z-boson does *not* have V−A couplings with the fermions. It must be kept in mind that when coupling it to Z_μ, this current should be multiplied by $g_z = \dfrac{g_2}{\cos\theta_W}$. Note that the expression of the current will remain the same even when it is written in terms of the mass eigenstates d^i of instead of d'^i.

The weak neutral current can also be written in terms of the T_3 and Y of the various fermions and also as a combination of V and A currents as follows.

$$\begin{aligned} J_\mu^{\rm Z} = \sum_f J_\mu^{\rm Z,f} &= \sum_f \left[\bar{f}\gamma_\mu f_L g_L^f + \bar{f}\gamma_\mu f_R g_R^f\right] \\ &= \left[\frac{1}{2}g_V^f \bar{f}\gamma_\mu f - \frac{1}{2}g_A^f \bar{f}\gamma_\mu\gamma_5 f\right]. \end{aligned} \tag{38}$$

Here the sum is over all fermions $f^i = u^i, d^i, e^i, \nu^i, i = 1-3$. The couplings $g_L^f, g_R^f, g_V^f, g_A^f$ can be read off from Eqs. 28 and 37 to be

$$\begin{aligned} g_L^f &= T_3(f_L) - \sin^2\theta_W\, Q_f, \quad g_V^f = T_3(f_L) + T_3(f_R) - 2\,Q_f \sin^2\theta_W \\ g_R^f &= T_3(f_R) - \sin^2\theta_W\, Q_f, \quad g_A^f = T_3(f_L) - T_3(f_R) \end{aligned} \tag{39}$$

In the above equation, we have written down $T_3(f_R)$ explicitly, which in the GSW model is zero, with a view to generalize the expressions for the weak neutral current, should the fermions belong to other representations of $SU(2)_L \times U(1)_Y$, other than the one in the GSW model. Recall that $Q_{\rm f}$ is the electromagnetic charge of the fermion in units of positron charge.

Note now that the form for the neutral current of Eq. 38 is exactly the same, for all the fermions of a given electrical charge and given values of the$SU(2)_L$ quantum numbers. Since in the GSW model, all the quarks or leptons of a *given electric charge and handedness* belong to the same representation of $SU(2)$ the weak neutral current automatically conserves 'flavour', be it the leptonic one or the quark one. This is indeed quite reassuring since the experiments had shown that while 'flavour' changing charged weak current (Eq. 28) exist, decays caused by 'flavour' changing weak neutral current, FCNC mentioned before, are either forbidden or suppressed by orders of magnitude. Their absence at the tree level is automatically guaranteed in the GSW model, just by the particle content. The values of $g_A^f, g_V^f, g_L^f, g_R^f$ for the fermions of the GSW model are given in the Table 5.

f	ν	e^-	u	d
g_L^f	$\frac{1}{2}$	$-\frac{1}{2}+\sin^2\theta_W$	$\frac{1}{2}-\frac{2}{3}\sin^2\theta_W$	$-\frac{1}{2}+\frac{1}{3}\sin^2\theta_W$
g_R^f	0	$\sin^2\theta_W$	$-\frac{2}{3}\sin^2\theta_W$	$\frac{1}{3}\sin^2\theta_W$
g_A^f	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$
g_V^f	$\frac{1}{2}$	$-\frac{1}{2}+2\sin^2\theta_W$	$\frac{1}{2}-\frac{4}{3}\sin^2\theta_W$	$-\frac{1}{2}+\frac{2}{3}\sin^2\theta_W$

Table 5: The values of axial and vector neutral current couplings g_A^f, g_V^f for the fermions of the GSW model. Also given are the neutral current couplings g_L^f, g_R^f for the left and right handed fermion fields.

Thus we see that in the GSW model, the weak neutral current couplings are completely determined by g_2 and $\sin\theta_W$. The weak neutral current involving ν^i is pure left handed just like the corresponding charged current, where as for the charged fermions the V-A mixture depends on the electromagnetic charge of the fermion because the relative weight of L and R currents is decided by the hypercharge Y. While the strength of the axial current is completely decided by the T_3 value of f_L^i, the vector coupling depends on the weak mixing angle θ_W. As we will see later, the experimentally determined value of $\sin^2\theta_W \sim 0.25$. As a result the weak neutral current coupling of the charged lepton (e, μ, τ) is in fact close to zero.

The interaction of all the quarks and leptons with the electroweak gauge bosons is encoded in the currents $J_\mu^{\rm em}$, $J_\mu^{\pm}$ and $J_\mu^{\rm Z}$ given by Equation31, first of Equations 28 and Eq.38. In low energy reactions, the appropriate way to adjudge the strength of processes mediated by the weak neutral current is to derive the current-current form of the interaction Lagrangian starting from Eq. 38. This is done by considering the matrix element of a four fermion scattering process and taking the limit in which the mass of the exchanged gauge boson is infinite. Let us consider the scattering process $f_1 + f_2 \to f_1 + f_2$ through the exchange of a massive $W^\pm$ (i.e., via charged current:CC) as indicated in the left panel of Fig. 8. The

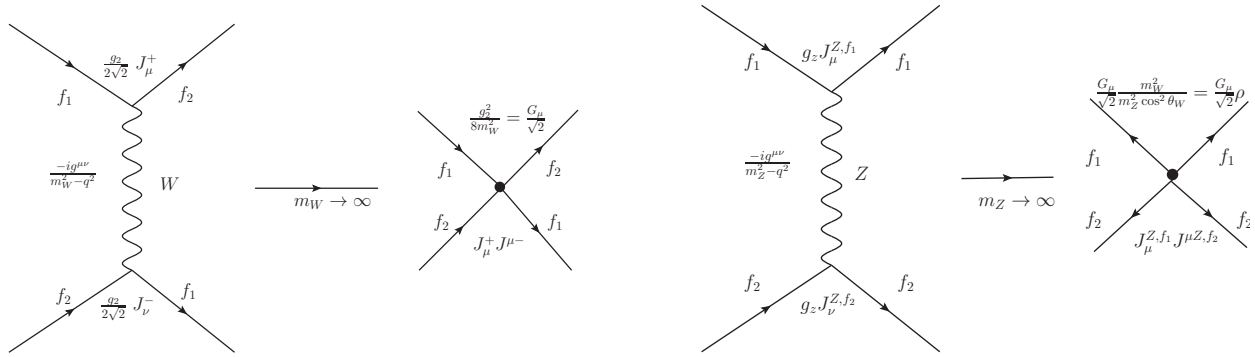


Fig. 8: Effective current current interactions for charged and neutral current processes in the left and right panel respectively.

effective current-current Lagrangian for the scattering process of Fig.8 can then be written as

$$\mathcal{L}_{\rm eff}^{\rm CC} = -\frac{g_2^2}{8M_W^2} J_\mu^+ J^{-\mu} = -\frac{G_\mu}{\sqrt{2}} J_\mu^+ J^{-\mu} \tag{40}$$

with $J_\mu^\pm$ as given by Eq. 28. On comparing with the current-current interactions of the pre gauge theory days, one then gets:

$$\frac{G_\mu}{\sqrt{2}} = \frac{g_2^2}{8M_W^2} = \frac{e^2}{8M_W^2 \sin^2\theta_W}, \tag{41}$$

where $G_\mu V_{ud} = G_F$. It can be noted here that since $|\sin\theta_W| < 1$, the experimentally measured value of G_μ and e, tells us that $M_W > 37.43$ GeV. For the limiting value of $\sin\theta_W \sim 1$ we get $M_W \sim 100$ GeV.

One can similarly write down the effective neutral current interaction effective Lagrangian under the approximation that the Z boson mass is large, by considering the four-fermion scattering process shown in the right panel of Fig. 8. This is given by

$$\mathcal{L}_{\rm eff}^{\rm NC} = -\frac{g_z^2}{2} \left(\sum_f J_\mu^{\rm Z,f} \right) \left(\sum_f J^{\mu,{\rm Z,f}} \right) \tag{42}$$

If one calculates the matrix elements for scattering process $\nu_e + e^- \to \nu_e + e^-$ taking place via the interaction of Eq. 40 and Eq. 42 respectively, *viz.*, $\mathcal{M}_{CC}$ and $\mathcal{M}_{NC}$, it can be seen that their ratio is given in terms of M_Z, M_W and $\sin\theta_W$ as:

$$\frac{\mathcal{M}_{NC}}{\mathcal{M}_{CC}} = \frac{M_W^2}{M_Z^2 \cos^2\theta_W} \equiv \rho. \tag{43}$$

Note further, that this effective Lagrangian involves couplings g_2, g_1 and M_W, M_Z. More directly we can use the two measured couplings G_μ and $\alpha_{\rm em}$ along with ρ and one arbitrary parameter of the model the weak mixing angle $\sin\theta_W$. M_W, M_Z are then given in terms of these and we have traded g_1, g_2 for G_μ and $\alpha_{\rm em}$. We will come back to this later in our discussion of the experimental validation of the SM.

Note also that in these discussions we have completely sidestepped the issue of how the non-zero masses for the gauge bosons and the fermions written can be made consistent with gauge invariance. In case of the gauge bosons the loss of gauge invariance also means loss of renormalisability and hence consequently of the ability to make any predictions. So one of the problems to be addressed is how to generate the mass terms below in a gauge invariant manner.

$$\mathcal{L}_{mass} = \frac{1}{2} M_Z^2 Z_\mu Z^\mu + M_W^2 W_\mu^+ W^{-\mu} + \sum_i m_i \left[\bar\psi_{iL}\psi_{iR} + \bar\psi_{iR}\psi_{iL} \right]. \tag{44}$$

It should be noted that the sum in Eq. 44 is over all the fermions except the neutrinos which are assumed to be massless here in this discussion.

3.2.2 SSB and generation of W/Z masses.

Before we move on to discuss more about the novel phenomenon of the existence of the weak neutral current, which was but the first step in testing and establishing the GSW model, let us first look at the issue of how nonzero masses for the gauge bosons and all the fermions can be generated in a gauge invariant manner. This is achieved [4] through the famous SSB mechanism [16].

One starts with the $SU(2)_L \times U(1)_Y$ gauge invariant Lagrangian, for the nonabelian gauge fields $W_\mu^i,\ i = 1, 3$ and the abelian gauge field B_μ, analogous to Eqs. 13 and 11 respectively.

$$\mathcal{L}_{massless} \quad = \quad \mathcal{L}_{gauge} + \mathcal{L}_{fermikin}$$

$$= \quad -\frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}F^a_{\mu\nu}F^{a,\mu\nu} + \mathcal{L}_{fermikin}.$$

Here $B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$ and $F^a_{\mu\nu} = \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g f^{abc} W^c_\mu W^c_\nu$ with $f^{abc} = \epsilon^{abc}$. Further, $\mathcal{L}_{fermkin}$ is given by Eq. 26.

The considerations of SSB begin by considering a complex scalar field Φ, which is a colour singlet and an $SU(2)_L$ doublet with hypercharge $Y_\phi = 1$, given by

$$\Phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} \equiv \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}$$

where $\phi_i = Re(\phi_i) + iIm(\phi_i)$ and similarly for ϕ^+, ϕ^0. Thus we have four real scalar fields and the Lagrangian we consider is,

$$\mathcal{L}_\Phi = (D_\mu\Phi)^\dagger D^\mu\Phi - V(\Phi) = (D_\mu\Phi)^\dagger(D_\mu\Phi) + \mu^2\Phi^\dagger\Phi - \lambda(\Phi^\dagger\Phi)^2, \tag{45}$$

with $\mu^2 > 0$. Note that compared to the Lagrangian for a free complex scalar field, this has the wrong sign for the quadratic term. So μ is not the mass and we can not interpret the excitations of the field Φ as propagating degrees of freedom. But it is precisely this wrong sign that is required for the spontaneous symmetry breaking to occur.

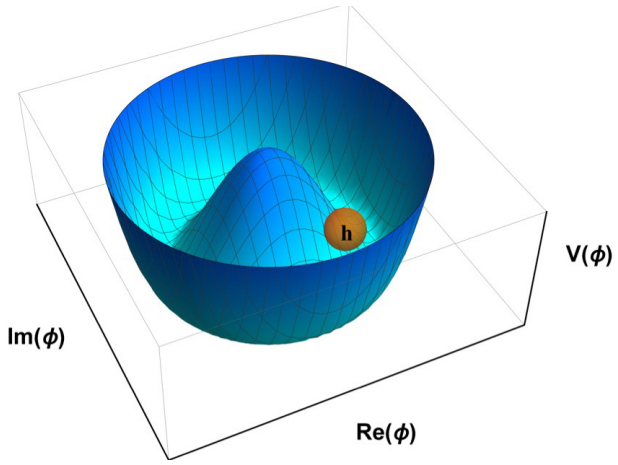


Fig. 9: A sketch of the mexican hat potential

Let us look at Figure 9 which shows a sketch of a similar potential, but for a single complex scalar field ϕ: $V(\phi) = -\mu^2\phi^\dagger\phi + \lambda|\phi^\dagger\phi|^2$. This shows clearly that classically the point $\Re e\phi = \Im m\phi = 0$ is in fact a maximum and there exist a continuum of minima where the field is nonzero, all related to each other by the symmetry transformation of the Lagrangian, which is a $U(1)$ transformation for the case shown in Fig. 9. SSB occurs when the quantum field configuration is such that the field has a nonzero vacuum expectation value corresponding to one of these minima, thus breaking the symmetry. The system is then described by the fluctuations of the fields around this minimum.

For the $V(\Phi)$ of Eq. 45 the minimum occurs for

$$\Phi^\dagger\Phi = \frac{\mu^2}{2\lambda} \equiv \frac{v^2}{2}. \tag{46}$$

The $SU(2)$ symmetry is broken when the vacuum field configuration chooses a particular direction in the ϕ^1, ϕ^2 space. The choice of the representation of the Higgs field decides pattern of symmetry breaking. For the case of $SU(2)_L \times U(1)_Y$ case under consideration, the unbroken symmetry should correspond to

the $U(1)_{\rm em}$ invariance since the γ is massless. Glashow's partial symmetry breaking with $Q = T_{3L}+Y/2$ aids in deciding how to implement and helps us decide which of the four scalar fields can acquire a nonzero vev. The charge operator should annihilate the vacuum and hence only the electrically neutral, real scalar field can have a nonzero vev. The required symmetry breaking pattern is guaranteed (with the choice $Y_\Phi = 1$) by

$$< 0|\Phi|0 >=< \Phi >_0= \begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix} \tag{47}$$

As follows from Eq. 46, $v = \sqrt{\frac{\mu^2}{\lambda}}$. Since Φ is a $SU(2)_L$ doublet clearly this choice for the vev means that the vacuum configuration breaks the symmetry and chooses a particular minimum from amongst the continuum of minima, similar to the situation depicted in the picture in Fig. 9. Since the electromagnetic charge still annihilates the vacuum, the symmetry breaking pattern is $SU(2)_L \times U(1)_Y \rightarrow U(1)_{em}$

One can rewrite the field Φ using the following parameterisation in terms of $\theta_a, a = 1, 3$ and h all of which have vacuum expectation value to be 0.

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} \theta_2 + i\theta_1 \\ v + h(x) - i\theta_3 \end{pmatrix}. \tag{48}$$

If $\theta_a(x), h(x)$ are small then we get

$$\Phi(x) = \exp(i\theta_a\tau^a/v) \begin{pmatrix} 0 \\ v/\sqrt{2} + h(x)/\sqrt{2} \end{pmatrix}. \tag{49}$$

This is then an expansion of the field Φ in terms of the fluctuations around the minimum. One recognizes the factor outside as that for a gauge transformation for a $SU(2)_L$ doublet. Comparing this expression with Eq. 16 we see immediately that by doing a gauge transformation $\Phi' = -\exp(i\theta_a\tau^a/v)\Phi$ we get,

$$\Phi'(x) = \begin{pmatrix} 0 \\ v/\sqrt{2} + h(x)/\sqrt{2} \end{pmatrix} \tag{50}$$

This gauge is called the Unitary gauge. Equation 47 also means that the vev is zero for field h. The three scalar degrees of freedom θ_i in fact have disappeared from the spectrum in this gauge. Indeed these three correspond to three Goldstone Bosons corresponding to the three generators of the symmetry group that are broken spontaneously.

Let us now evaluate $\mathcal{L}_\Phi$ of Eq. 45 in the unitary gauge using Φ' from Eq. 50. We use

$$D_\mu\Phi = \partial_\mu\Phi - i\frac{g_1}{2}B_\mu\Phi - ig_2W_\mu^a\frac{\tau^a}{2}\Phi. \tag{51}$$

The covariant derivative term in Eq. 45 gives rise to terms quadratic in the gauge boson fields which are given as below:

$$\begin{aligned} \left|\left(\frac{g_1}{2}B_\mu + g_2\frac{\tau^a}{2}W_\mu^a\right)\begin{pmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{pmatrix}\right|^2 &= \frac{g_2^2v^2}{8}\Big(W_\mu^aW^{a\,\mu}\Big) + \frac{g_1^2v^2}{8}B_\mu B^\mu \\ &\quad -\frac{g_1g_2v^2}{4}W_\mu^3B^\mu \\ &= \frac{g_2^2v^2}{4}W_\mu^+W^{-\,\mu} + \frac{v^2}{8}\Big(g_1B_\mu - g_2W_\mu^3\Big)^2 \\ &= \frac{g_2^2v^2}{4}W_\mu^+W^{-\,\mu} + \frac{(g_1^2+g_2^2)v^2}{8}Z_\mu Z^\mu \end{aligned} \tag{52}$$

This then tells us directly that three of the four gauge bosons become massive: the $W^{\pm}$ and one linear combination of B_μ, W^3_μ which we call Z_μ and the orthogonal linear combination remains massless. This also tells us

$$M_W^2 = \frac{g_2^2 v^2}{4}, \quad M_Z^2 = \frac{(g_1^2 + g_2^2)v^2}{4} = \frac{M_W^2}{\cos^2\theta_W}. \tag{53}$$

Identifying $g_1 B_\mu - g_2 W^3_\mu$ with Z_μ with proper normalisation we see that expression for Z_μ is the same as that given in Eq. 22 and $\tan\theta_W$ same as that in Eq. 33.

The new thing compared to the earlier discussion of the GSW model, is that now one has a model for generating masses for the gauge bosons from the gauge invariant kinetic term of the scalar field. The combination A_μ remains massless as it must. The fact that the same linear combination which has mass zero also has the couplings to fermions that a photon field A_μ must have (cf. Eqs. 30,31) means that the SSB has achieved the desired symmetry breaking pattern. Further, in the earlier discussion M_W, M_Z were unknowns, put in by hand; but now we find that the two are related to each other.

Another fact worth noticing is that the value of the vev v gets determined in terms of measured value of G_μ. Using the expression for M_W in Eq. 53 and that for G_μ in Eq. 41, we get

$$v = \left(\frac{1}{\sqrt{2}G_\mu}\right)^{1/2} \simeq 246 \text{ GeV}. \tag{54}$$

Using the expression for g_2 in terms of e and $\sin\theta_W$ and Eq. 53, one can then see that,

$$M_W = \sqrt{\frac{\pi}{\sqrt{2}G_\mu}\frac{\alpha_{\rm em}}{\sin^2\theta_W}} = \frac{37.3}{\sin\theta_W} \text{ GeV}; \qquad M_Z = \frac{37.3}{\sin\theta_W\cos\theta_W} \text{ GeV}. \tag{55}$$

This is the promised reduction in the number of free parameters. Now everything in the GSW model is predicted in terms of the two known constants $\alpha_{\rm em}, G_\mu$ and one free parameter $\sin^2\theta_W$. An accurate determination of G_μ is possible via life time of the muon, τ_μ. Since $|\sin\theta_W| < 1$ this also means we have an automatic lower limit on the masses of the W, Z bosons of 37.3 GeV.

We further notice from Eq. 53 that the ratio ρ defined in Eq. 43 is *predicted* to be unity in the GSW model and we have

$$\rho = \frac{M_W^2}{M_Z^2\cos^2\theta_W} = 1.$$

Noting, in addition, from Table 5 that g_A^f, g_A^f are numbers of $\mathcal{O}(1)$, we can then conclude from Eqs. 40–42 that one should expect the ν induced scattering processes via neutral current interactions to happen at rates similar to those via charged current interactions. This conclusion is of course independent of the actual values of M_W, M_Z with the proviso that the energies are much smaller compared to these masses. Thus the GSW model not only predicted the existence of a weak neutral gauge boson and weak neutral current processes mediated by it, but it also predicted their strength to be $\mathcal{O}(G_\mu)$.

The experiments with the bubble chamber Gargamelle at CERN, found evidence for the processes induced by neutral current interactions as predicted by the GSW model. The energies involved were smaller than the lower bound on the W/Z masses implied by Eq. 55. Hence one can use the effective lagrangian description of Eqs. 40 and 42. In addition, measurements of cross-sections for neutral current processes further showed the ratio ρ to be close to 1. Thus these provided both the qualitative and quantitative support for the GSW model. This was before W, Z were experimentally discovered and their masses measured.

It was further seen that the model prediction of $\rho = 1$ is true even with additional Higgs fields as long as the scalars responsible for the SSB belong to the doublet representation. This can be understood in terms of an accidental symmetry that the scalar potential $V(\Phi)$ seems to have for this choice of the representation of the Higgs field. We shall discuss later this symmetry called the Custodial Symmetry.

After working out the remaining terms also in terms of the field Φ' in the unitary gauge we get,

$$\begin{aligned}\mathcal{L}^U_{\Phi'} &= \left[M_W^2 W^+_\mu W^{-\mu} + \frac{1}{2} M_Z^2 Z_\mu Z^\mu\right]\left(1+\frac{h}{v}\right)^2 + \frac{1}{2}(\partial_\mu h)^2 - \mu^2 h^2 - \lambda v h^3 - \frac{\lambda}{4}\, h^4 \\ &= \mathcal{L}_{VVh} + \mathcal{L}_h. \end{aligned} \tag{56}$$

The first two terms are the mass terms for the W, Z as well as the term describing the interaction between a pair of gauge bosons and the h. The form of this term makes it very clear that the strength of the VVh coupling is simply proportional to the mass of the corresponding gauge boson. This proportionality between the mass and the coupling is the most critical prediction of the SSB.

The remaining terms describe now a real, scalar field which is a propagating degree of freedom with mass $M_h = \sqrt{2\mu^2}$. Since $v = \sqrt{\mu^2/\lambda}$, the mass of the Higgs boson is given in terms of self coupling λ. This being an arbitrary parameter of the Higgs potential, not fixed by any condition, M_h too is a free parameter of the SM, with no prediction for it. We will come back to this later when we look at theoretical constraints on the Higgs mass!

In the unitary gauge now the propagating degrees of freedom are the three massive gauge bosons $W^\pm, Z$, one massless gauge boson γ and ONE propagating massive scalar. A massless vector boson has two degrees of freedom corresponding to the two degrees of polarisation it can have whereas a massive gauge boson has three degrees of freedom as it can also have longitudinal polarisation. Out of the four scalar degrees of freedom only one, h, is left in the particle spectrum and the other three provide the remaining degrees of freedom corresponding to the longitudinal polarisation necessary for the three gauge bosons to be massive. The total number of bosonic degrees of freedom before SSB are twelve: eight corresponding to four massless gauge boson fields $W_\mu^{a,a=1,3}, B_\mu$ and the four scalars in Φ. After the SSB one has again twelve bosonic degrees of freedom : nine corresponding to the three massive gauge bosons $W^\pm, Z$, two corresponding to the massless photon γ and one corresponding to the massive neutral scalar h. In the unitary gauge the particle spectrum contains only the physical fields and the Goldstone boson fields $\theta_a, a = 1, 3$ of Eq. 48, are absent from the spectrum. The same is depicted somewhat pictorially below:

$\mathcal{L}^{massless}_{gauge}$	+ $\mathcal{L}_\Phi$		$\mathcal{L}^{massive}_{gauge}$	+ $\mathcal{L}_h$
4 massless gauge bosons	4 scalar fields	$\xrightarrow{SSB, Unitary gauge}$	3 massive, 1 massless gauge bosons	1 physical scalar
8 d.o.f.	4 d.o.f.		11 d.o.f	1 d.o.f.

Table 6: Bosonic degrees of freedom before and after the SSB.

3.2.3 SSB and generation of lepton masses

It was really Weinberg's genius that he saw that exactly the same mechanism can be used effectively to give masses to *all* the fermions. He did so by postulating a gauge invariant term for interaction between the fermionic matter fields and the Higgs field! For the electron, it can be written as

$$\mathcal{L}^e_{yukawa} = -f^{*e}\bar{\mathcal{L}'}_{1L}\Phi e'_{1R} + h.c. \tag{57}$$

The 'prime' on the lepton fields are to indicate that these the interaction eigenstates. One can also see clearly that this is a singlet under $SU(2)_L$ and $U(1)_Y$). Using Φ' of Eq. 50, we get we get

$$\mathcal{L}^{e,U}_{yukawa} = -\frac{f^{*e}v}{\sqrt{2}}(\bar{e}'_L e'_R)(1 + h/v) + h.c. \tag{58}$$

The first term in the bracket is clearly the mass term. Hence we have

$$m_e = +f^{*e}v/\sqrt{2} \qquad\qquad e' = e \tag{59}$$

Second term in the bracket also then tells us that the hee coupling is just m_e. One can do the same for all the charged leptons. Thus the *gauge invariant* Lagrangian $\mathcal{L}^i_{yukawa}$, gives rise to the mass term for the leptons.

The original paper by Weinberg [11] talked *only* of leptons. With some extra work the procedure works for the case of quarks as well. The most general Yukawa interaction can be written as,

$$\mathcal{L}^q_{yukawa} = -f^{*d}_{ij}\bar{\mathcal{Q}}'^i_L \Phi d'i_R - f^{*u}_{ij}\bar{\mathcal{Q}}'^i_L \tilde{\Phi} u'^j_R + h.c. \tag{60}$$

where $\tilde{\Phi} = i\sigma_2\Phi^*$. We want the $\mathcal{L}$ to be invariant under $SU(2)_L \times U(1)_Y$ transformations. The $SU(2)_L$ invariance is guaranteed by construction. Recall, for the right handed quark fields the hyper charges are $Y = -\frac{2}{3}$ and $\frac{4}{3}$ for the down-type and up-type quarks respectively whereas $\bar{\mathcal{Q}}^{i'}$ has $Y = -\frac{1}{3}$. As a result, the second term involving up-type quarks in $\mathcal{L}^q_{yukawa}$ is invariant *ONLY* if the hypercharge of the scalar doublet has $Y = -1$. The most economical choice for such a field is then $\tilde{\Phi}$. Again the $'$ for the quark fields indicate that these are interaction eigenstates. In the unitary gauge, using Φ' of Eq. 50 we get,

$$\mathcal{L}^{q,U}_{yukawa} = -\frac{f^{*d}_{ij}}{\sqrt{2}}v\ \bar{d}'^i_L(1 + h/v)d'^j_{jR} - \frac{f^{*u}_{ij}}{\sqrt{2}}v\ \bar{u}'^i_L(1 + h/v)u'^j_R + h.c. \tag{61}$$

We see that after the SSB, the $SU(2)_L \times U(1)_Y$ gauge invariant Lagrangian $\mathcal{L}^q_{yukawa}$ of Eq. 60 contains mass terms for both the up-type and down-type quarks. These are matrices in the generation space and are given by;

$$m^d_{ij} = \frac{f^{*d}_{ij}}{\sqrt{2}}v \qquad , \qquad m^u_{ij} = \frac{f^{*u}_{ij}}{\sqrt{2}}v. \tag{62}$$

Since in general f^{*d}_{ij}, f^{*u}_{ij} are completely arbitrary matrices in the generation space, these mass matrices are not diagonal in the basis d'^i, u'^i, in the most general case. The states $d'^i, u'^i, i = 1-3$ are therefore clearly not mass eigenstates. $d^i, u^i, i = 1, 3$ are thus linear combinations of $d'^i, u'^i, i = 1, 3$. In the most general case, after diagonalisation of both the m^d, m^u matrices given above, we can write the weak charged current in terms of the mass eigenstates u^i, d^i as indicated in Eqs. 27 and 28. An alert reader might have wondered why one does not have such a mixing matrices for the charged leptons. This has to do with the fact that the mixing matrix $\mathbf{V}$ given in Eq. 6, arises from a mismatch in the matrices which diagonalise the d and u mass matrices, and will be different from each other in the most general case. However, for the charged lepton case, the neutrinos being massless, the corresponding mismatch between matrices diagonalising the charged lepton and neutrino mass matrices, can not have any physical implications.

3.2.4 Flavour changing neutral currents

An alert reader might wonder why we emphasize the issue of FCNC so much. To appreciate this, we have to discuss briefly one more puzzle that the weak decays of the K mesons had presented to the theorists during the development of a theory of weak interactions. Let us consider the leptonic decay of $K^+ \to l^+\nu_l$. The big difference in the measured branching ratios for the leptonic decays $l\nu_l$, $(63.55 \pm 0.11)\%$ and $(1.581 \pm 0.007) \times 10^{-5}$ for $l = \mu, e$ respectively, can be understood in terms of the $V - A$ structure of the leptonic current in first of the equations in Eq. 28. The $K^\pm$ were known to have a non-leptonic decay as well, with a branching ratio of about 25%. On the other hand, the K^0 mesons were found to decay only in the non-leptonic final states. For example, even today only an upper limit of 9×10^{-9} is available for the branching ratio for $K^0_S \to \mu^+\mu^-$, meaning thereby that this decay is not yet seen. This big difference in the leptonic branching ratios for the $K^\pm$ on the one hand and K^0_S on the other, was

interpreted as suppression of strangeness changing weak neutral current as compared to the strangeness changing, weak charged current. However, there was no 'understanding' as to why this should be so. So after the postulation of weak neutral currents in the GSW model, it was an obvious question to ask whether the model provides a 'natural' understanding of the observed fact of suppression of the flavour changing weak neutral currents.

Weak decays of hadrons can be understood (and calculated) in the framework of the quark model and $W^{\pm}$ bosons. The left panel of Fig. 10 shows the diagram which needs to be computed for (say) the $\Delta S = 1$ weak decay, $K^+ \to \nu_l l^+$ taking place via charged current. The hadronic decays of the

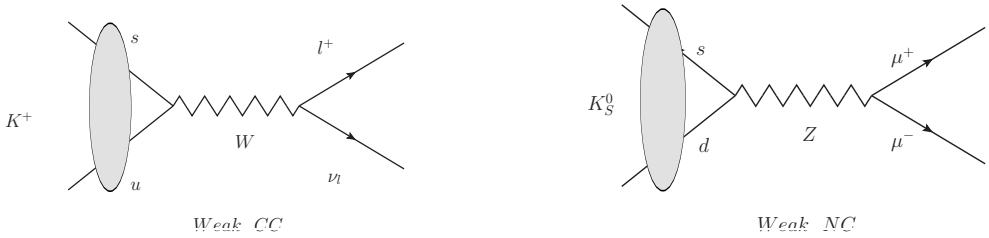


Fig. 10: Leptonic decay of mesons via currents. The blob indicates that the quarks are bound in the K mesons.

$K^{\pm}$ mesons can then be understood in terms of hadronic decays of the $W^{\pm}$. Both the non-leptonic and leptonic decays of the $K^{\pm}$ thus happen at the weak rate; amplitude being proportional to G_{μ}, the relative branching ratios being controlled by those of the $W^{\pm}$ which are known in the GSW model.

The existence of the weak neutral Z boson, in principle, could have given rise to weak leptonic decay of K_S^0 mediated by the Z as depicted in the right hand panel of Fig. 10 with rates similar to the charged weak current processes, should a $u - d - Z$ vertex exist. This too would be then a $\Delta S = 1$ process. The happy instance of absence of such a term in the J_{μ}^Z of Eq. 38, explains the absence of pure leptonic decays of K_S^0 via the weak neutral current at the tree level. This is then consistent with the experimentally observed suppression of such decays. As has been already mentioned, absence of this current is due to the fact that the fermions of the SM with a given electromagnetic charge and handedness, belong to the *same* representation of the EW gauge group. Thus, the observed suppression of the FCNC decays, in fact indicated the *need* of the existence of the c quark with $Q = \frac{2}{3}$, which is a $T_3 = \frac{1}{2}$ member of the $SU(2)$ doublet along with the s quark. The mere presence of a c-quark in the spectrum is enough to achieve this absence of the FCN. Further, this result is independent of the masses of the quarks involved.

Even though such a decay is forbidden at the *tree level* by the absence of FCNC couplings in Eq. 38, it can take place through loop processes at a higher order in G_{μ} through the charged current (CC) interactions. In a renormalisable gauge theory such as the GSW model, one should be able to compute the rate at which it is predicted to occur. This can then be compared with the observed suppression of less than one part in 10^9.

Fig. 11 depicts two of the possible four box diagrams which would give rise to this decay at the loop level, in a world with only four quarks u, d, s and c. The difference between the left and the right panel is in identity of the charge $+\frac{2}{3}$ quark which is exchanged in the t-channel. There will also be two more diagrams where the W's form the vertical legs of the box. One calculates these loops explicitly in a gauge theory with SSB as it is renormalisable. In a world with only 3 quarks, one would compute only the diagram in the left panel where the u quark is exchanged in the t channel and the amplitude of a second diagram where it is W which is exchanged in the t channel and u forms the horizontal leg of the box. Recall we already know that for a unified theory $M_W > 37.3$ GeV. The loop amplitude, can then be computed in the approximation $m_u^2 << M_W^2$. The amplitude of the box will be proportional to $G_{\mu}^2 \sin\theta_c \cos\theta_c \times \text{loop function}$, modulo the wave function factors which will describe how the $\bar{s}$ and

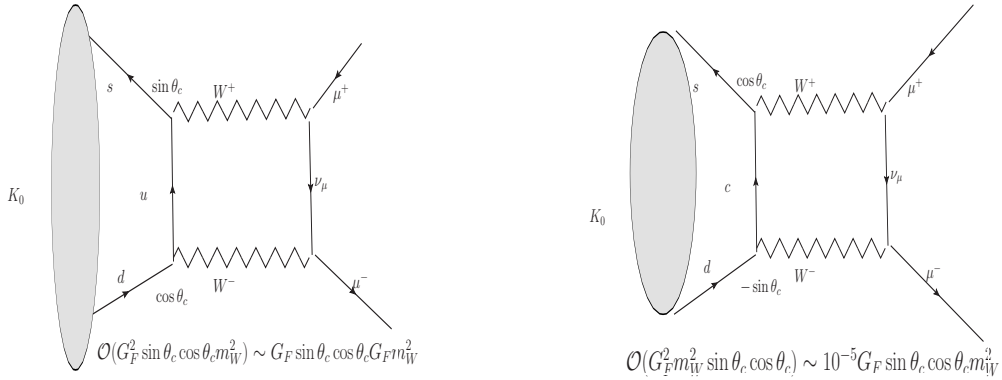


Fig. 11: One loop diagrams giving rise to $K_S^0 \to \mu^+\mu^-$

d quarks are held together to form a $\bar{K}_S^0$. One then gets

$$\mathcal{M}_{\mu\mu}^{loop}(K^0 \to \mu^+\mu^-) \propto \frac{g_2^2}{M_W^4} \cos\theta_c \sin\theta_c \;\; g_2^2 \times M_W^2 (1 + \mathcal{O}(m_u^2/M_W^2) \tag{63}$$

The factors of $\sin\theta_c, \cos\theta_c$ that appear at various vertices in these diagrams are a reflection of Cabibo mixing. In the limit where all the masses can be neglected, the loop function can only involve M_W^2, which is what explicit computations will yield. The M_W^4 in the denominator comes from the W-propagators. Remembering the relation between G_μ and M_W^2 (Eq. 41), we then find that the amplitude can be written as:

$$\mathcal{M}_{\mu\mu}^{loop}(K^0 \to \mu^+\mu^-) \sim G_\mu^2 \cos\theta_c \sin\theta_c M_W^2. \tag{64}$$

Let us compare then the order of magnitude for this amplitude with the one expected for the non leptonic

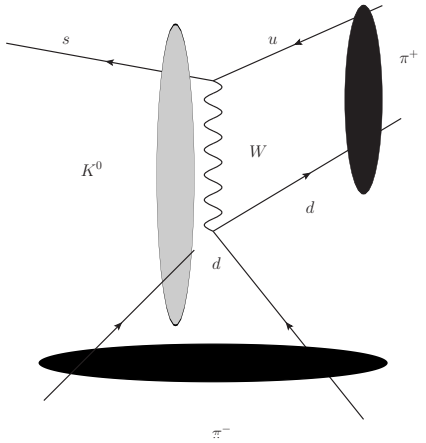


Fig. 12: One loop diagrams giving rise to $K_S^0 \to \pi^0\pi^0$

weak decay $K_S^0 \to \pi^0\pi^0$. The latter takes place not through a loop diagram but via the weak charged current at tree level and occurs at $\mathcal{O}(G_\mu)$. A possible digram is shown in Fig. 12. Amplitude for this decay will be proportional to $G_\mu \sin\theta_c \cos\theta_c$, modulo the aforementioned wave function factors describing $q\bar{q}'$ bound state. If it were not for the factor of M_W^2 ($M_W > 37.3$ GeV^2), the additional factor of G_μ present in the loop amplitude of Eq. 64, could have suppressed the $\mathcal{M}_{\mu\mu}^{loop}$ by a factor 10^{-5} compared to the charged current induced, tree level amplitude for $K^0 \to \pi^0\pi^0$. Thus the rate for the $\mu^+\mu^-$ decay could have been suppressed to the experimentally observed low level as compared to the $\pi^0\pi^0$ decay.

However, the factor M_W^2 removes this suppression of the $\mathcal{M}_{\mu\mu}^{loop}(K^0 \to \mu^+\mu^-)$. As a result, in the three quark picture, the amplitude for the $\mu^+\mu^-$ decay is suppressed though not *hugely* compared to the $\pi^0\pi^0$ decay which in turn occurs at the usual weak rate. This then is in contradiction with the experimentally observed branching ratio of about $\sim 31\%$ for the $\pi^0\pi^0$ final state and the observed upper limit on the branching ratio for the $\mu^+\mu^-$ channel of 10^{-9}.

When one adds to the loop amplitude of Eq. 63 the contribution coming from c loop as well, something interesting happens. Due to the relative negative sign of the term containing $\sin\theta_c$, we note that the amplitudes from the two box diagrams in the left and right panel of Fig. 11, will cancel each other exactly in the case where the masses of the u and c quarks are equal. The large term independent of the mass of the quark in the loop thus cancels between these two diagrams! The non leading terms dependent on the mass of the quark in the loop, will give zero when $m_u = m_c$ and will be proportional to $m_c^2 - m_u^2$. So the factor with mass dimension two, in the amplitude $\mathcal{M}_{\mu\mu}$ is no longer the large M_W^2, but $m_c^2 - m_u^2 \sim m_c^2$. Thus, in the four quark picture, the observed suppression happens due to the very existence of the charm quark and is guaranteed here by the *orthogonality* of the quark mixing matrix. Further, any deviation from zero for the branching ratio will then depend on the difference in the masses of the quarks being exchanged in the loops and in fact can give *indirect* information on these, in the framework of a gauge theory when the various parameter values g_1, g_2, v and mixing angles are known. However, particularly in the case of $K^0 \to \mu^+\mu^-$ no firm constraint on the charm mass can be drawn due to the existence of additional contributions to this process which do not come from the weak charged current interactions along with some accidental cancellations.

A similar suppression of FCNC is also observed experimentally in the the K^0-$\bar{K}^0$ mixing which is a $\Delta S = 2$ transition. In principle, this could occur at higher order in the CC weak interactions which are strangeness changing with $\Delta S = 1$. The K_L–K_S mass difference is $\Delta m_K = |m_{K_L} - m_{K_S}| = (3.484 \pm 0.006) \times 10^{-12}$ MeV, with $\frac{\Delta m_K}{m_{K^0}} \simeq 8.5 \times 10^{-15}$. Recall here that the strength of weak interactions is given by $G_\mu \sim \frac{1.01\times10^{-5}}{m_p^2}$. The strength of the $\Delta S = 2$ transition which causes the K^0–$\bar{K}^0$ oscillations and gives rise to the K_L–K_S mass difference, is thus clearly weaker than that expected from just two insertions of the CC weak interaction and is thus suppressed perhaps even further. In the early days of gauge theory it was not clear whether the K_0–$\bar{K}_0$ mixing is caused by a new interaction *weaker* than the weak or whether it can be understood as a higher order effect of the $|\Delta S| = 1$ weak charged current interaction.

In a gauge theory one can compute the expected value of this mixing in terms of loop diagrams very similar to those shown in Fig. 11, where at the right hand end of the box the ν_μ is replaced by a u or c-quark line and the μ^+, μ^- lines are replaced by the $\bar{d}$ and s quark line which are bound in a $\bar{K}^0$ meson. Again, we show only two of these diagrams contributing to it and that too in the 4-quark picture, in Fig. 13. Again one can make very similar observations as before. If the model had only three quarks u, d, s then only the digram involving the u quarks would have contributed and it is very clear that the predicted K_0–$\bar{K}_0$ mass difference will not be proportional to G_μ^2 in the limit that u, d, s quark masses are much smaller than M_W. As a result this contribution would have been much bigger than the experimental measurement mentioned above. On the other hand, in the four quark picture, if the masses of u and c quarks were equal the contribution from the two diagrams will just cancel each other due to the factors of $\cos\theta_c$ and $\sin\theta_c$ appearing with appropriate signs and will be zero in this limit of $m_c = m_u$. Further, the actual value of the predicted mass difference will now depend on m_c, m_u as well as experimentally measured values of G_μ, θ_c etc. The observed mass difference could then be interpreted as an upper limit on the mass difference $m_c - m_u$ and further as a limit on m_c of about a few GeV neglecting m_u. This is perhaps the first example of prediction of the 'scale' of new physics (in this case the charm quark) through virtual effects on quantities measured at energies much below the scale.

There are two parts to this calculation. One is evaluation of the transition amplitude indicated by the box diagram drawn involving the W's and the quark lines, and the other is conversion of that

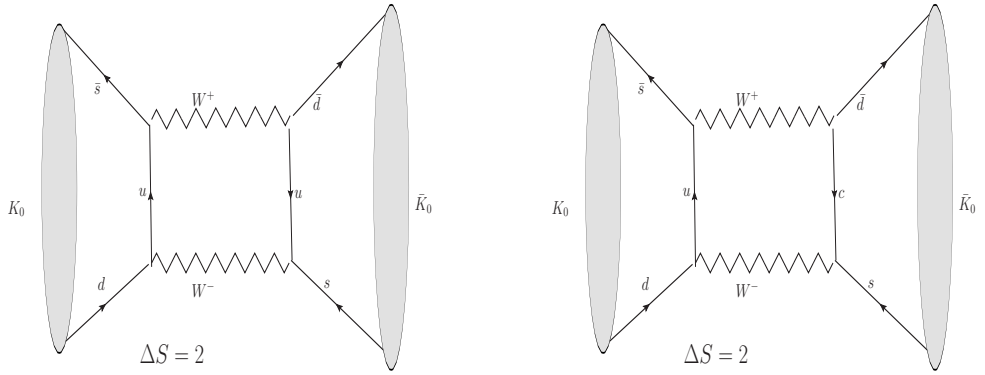


Fig. 13: One loop diagrams giving rise to $K^0 - \bar{K}^0$ mixing.

amplitude into mass difference between the mesons. This requires evaluation of the matrix element between the meson states of the effective Lagrangian which in turn has been extracted from the transition amplitude at the quark level. One can relate the former to the meson wave function factor encoded in the decay constant f_K which in turn can be extracted from the measured life times of the kaons. The loop calculation, yields a result for the mass difference $\Delta M_K = |M_{K_0} - M_{\bar{K}_0}|$,

$$\frac{\Delta M_K}{M_K} = \frac{2}{3}\frac{G_\mu^2}{4\pi^2} m_c^2 \cos^2\theta_c \sin^2\theta_c f_K^2 \tag{65}$$

In principle, the large mass of the t quark means that this could change substantially in the six-quark picture. A calculation of the mass difference in the six-quark case can be shown to be

$$\frac{\Delta M_K}{M_K} = \frac{2}{3}\frac{G_\mu^2}{4\pi^2} m_c^2 \cos^2\theta_c \sin^2\theta_c f_K^2 X \tag{66}$$

with

$$X = (\sin^2\theta_c \cos^2\theta_c)^{-1} \Re e \left[(V_{cs}V_{cd}^*)^2 + \frac{m_t^2}{m_c^2}(V_{ts}V_{td}^*)^2 + V_{cs}V_{cd}^* V_{ts}V_{td}^* \frac{2m_c^2}{m_t^2 - m_c^2} \ln\left(\frac{m_c^2}{m_t^2}\right) \right] \tag{67}$$

For the four-quark case the CKM matrix is just a 2×2 matrix and hence Eq. 66 just reduces to Eq. 65. For the six quark case, indeed X in Eq. 67 contains terms $\propto m_t^2$. These terms can, in principle, dominate the mass difference ΔM_K. However, since the elements of the CKM matrix which connect the third generation with the first and the second generation, V_{td}, V_{ts}, are extremely small, the dominant contribution to $\frac{\Delta M_K}{M_K}$ is still given by Eq. 65.

In fact, even without calculating the loop one could try to estimate the size of expected value of Δm_K assuming that the $\Delta S = 2$ transitions are caused by an interaction with strength proportional to $G_K^2 = G_\mu^2 \sin^2\theta_c$. Since G_μ has mass dimension -2, we need to add appropriate factors of the only mass available at the meson level, viz. m_K. Thus the expected mass difference is

$$\frac{\Delta M_K}{M_K} = G_\mu^2 \times m_K^4 = (1.01 \times 10^{-5})^2 \times \left(\frac{m_K}{m_p}\right)^4 \sin^2\theta_C \simeq \mathcal{O}(10^{-14}) \tag{68}$$

which is indeed the right order of magnitude. This thus means that this amplitude must be $\propto G_\mu^2 \sin^2\theta_c$ and can NOT be $\sim \mathcal{O}(G_\mu)$.

Thus one sees that the suppression of FCNC that has been observed experimentally is 'understood' neatly, both at the tree and loop level in a gauge theory, in terms of the chosen particle spectrum of the

SM. At the tree level case it is just guaranteed by the representation of the group to which quarks of a given electromagnetic charge and handedness belong where as at the loop level it is the orthogonality of the mixing matrix. I.e, the mere presence of charm quark in the spectrum is sufficient to achieve both. The latter observation is the celebrated GIM mechanism [17]. In the six quark case, it is not the orthogonality of the mixing matrix but the Unitarity of $\mathbf{V}$ matrix that guarantees the GIM cancellation. Further, the actual observed suppression can give a hint about the masses of the quarks involved. In fact, the first 'prediction' [18] for the charm mass around a scale $\lesssim$ a few GeV was made, using the GIM idea by comparing the observed ΔM, with the one calculated theoretically. The uncertainties in the upper limit were mainly due to the gaps in the theoretical understanding of strong interactions at the time. As explained above, while in principle this 'prediction' could have had 'large' corrections, for the values of the mixing matrix elements realised in nature, the prediction was correct.

3.2.5 Anomaly cancellation

As we have seen above, the GSW model contains both the vector and the axial vector currents. This causes a problem when we try to renormalise the theory and do loop computations. The gauge invariance of axial vector currents of the type

$$J^5_\mu = \bar{\psi}\gamma_\mu\gamma_5\psi',$$

($\psi' = \psi$ for neutral currents) is not preserved by dimensional regularization due to the presence of γ_5 in the current. This means that even though,

$$\partial_\mu J^\mu_5 = 0$$

classically, at loop level due to the non invariance of the regulator,$\partial_\mu J^\mu_5 \neq 0$ and the RHS develops a nonzero term on the RHS. Hence, this axial gauge current is no longer conserved. The current is said to be 'anomalous'. As we know from Noether's theorem if the current is not conserved, it means gauge invariance is broken. Gauge symmetry along with Higgs mechanism is needed to have a consistent quantum theory with massive gauge bosons. Thus if the theory has an anomalous current (or has anomaly) the theory may not make sense at quantum level. It was shown by Adler and Bell-Jakciw, that there is only one type of loop diagram with a logarithmic divergence which can make $\partial_\mu J^\mu_5$ non- vanishing and poses a danger to the conservation of the axial gauge current. This is a triangle diagram with a fermion loop and two gauge boson legs and one current insertion; equivalently one can also consider a fermion loop with three gauge boson legs. In the GSW model with its $SU(2)_L$ gauge bosons which have couplings only to left chiral fermions and the $U(1)_Y$ gauge bosons which have unequal couplings to the left and right chiral fermions, these triangle diagrams are in general not zero. Further, one can show that the anomalous contribution is independent of the mass of the fermions in the internal loop.

There are in fact four types of triangle diagrams we need to consider out of which three are shown in Fig. 14. Consider the diagram in the left most panel which contains matrix element of a pure $V - A$

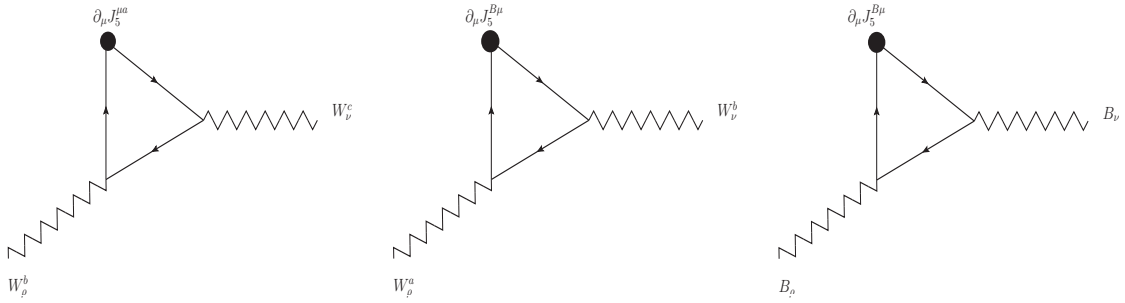


Fig. 14: Triangle diagrams with anomalies.

current insertion along with two $SU(2)_L$ gauge boson legs. Only left handed fermions contribute to this

anomaly and it can be shown that

$$\partial_\mu J_5^{\mu a} \sim tr\tau^a\{\tau^b,\tau^c\}\epsilon^{\alpha\rho\beta\nu}F^b_{\alpha\rho}F^c_{\beta\nu}. \tag{69}$$

Here the 'tr' refers to the trace over representation matrices and indicates the sum over all the fermions in the representation. Since $\{\tau^b,\tau^c\} = 2\delta^{bc}$ and τ^a are traceless matrices this anomaly is zero identically. In fact, the diagram with just one $SU(2)_L$ V–A current insertion not shown here will also give zero contribution to the anomaly due to the traceless property of $\tau^a, a = 1, 3$ matrices. The central diagram also gets contribution only from the left chiral fermions and is given by

$$\partial_\mu J_5^{\mu} \sim tr(Y_L)\epsilon^{\alpha\rho\beta\nu}F^a_{\alpha\rho}F^a_{\beta\nu} \tag{70}$$

The notation $tr(Y_L)$ indicates that only the left chiral fermions contribute to this quantity and sum is to be taken over one $SU(2)_L$ representation. The contribution of the rightmost diagram in Fig. 14 is given by

$$\partial_\mu J_5^{\mu} \sim tr\left(Y_L^3 - Y_R^3\right)\epsilon^{\alpha\rho\beta\nu}B_{\alpha\rho}B_{\beta\nu}. \tag{71}$$

We see that for a single lepton generation the anomaly of Eq. 70 is proportional to $2 \times Y_L = -2$. Summing over all the lepton doublets it will have a value -6. However, one notices that, for a single quark generation it is $2 \times 1/3$. The three colours add another factor of 3. Thus we find,

$$tr(Y_L)|_l + tr(Y_L)|_q = -2 + 3 \times 2 \times 1/3 = 0.$$

Thus this anomaly vanishes identically for the particle content of the left chiral fermions in the GSW model. Further, we also notice that while $(2Y_L^l)^3 - (Y_R^e)^3 = -2 + 8 = 6$ is not zero, it is again compensated by the value for the quark doublets which is $3 \times (-2/27 - \left(\frac{4}{3}\right)^3 + \left(\frac{2}{3}\right)^3) = -6$. Thus again

$$tr\left(Y_L^3 - Y_R^3\right)|_l + tr\left(Y_L^3 - Y_R^3\right)_q = 6 - 6 = 0.$$

Hence contributions to both the anomalies, from loops of fermions of one quark and one lepton doublet of the GSW model, are equal and opposite in sign. This means that the numbers of the lepton and quark doublets have to be exactly equal so that the anomalies do not spoil the gauge invariance of the GSW model and hence the renormalisability.

3.2.6 Custodial Symmetry

Let us discuss further the ρ parameter. To that end let us understand in a little more detail the origin of the prediction of unity for ρ defined Eq. 43. Let us begin first writing down the most general gauge boson mass terms that one could generate by spontaneous symmetry breaking. In the $W^a_\mu, a = 1, 3$ and B_μ basis this can be written as

$$\begin{pmatrix} M^2_{W^1} & 0 & 0 & 0 \\ 0 & M^2_{W^2} & 0 & 0 \\ 0 & 0 & M^2_{W^3} & M_{WB} \\ 0 & 0 & M_{WB} & M^2_B \end{pmatrix} \tag{72}$$

The mass terms $M_{W^a}, a = 1 - 3, M_B$ and M_{W^3B} arise from the covariant derivative term $D_\mu\Phi^\dagger D^\mu\Phi$ (cf. Eq. 51), after the field Φ acquires a non zero vev. The expressions for M_{W^3B}, $M_{W^a}, a = 1, 3$ and M_B that one would get as a result by expanding the field around the minimum, will depend on the weak isospin charges T, T_{3L} of the field Φ. Demanding that the EW minimum conserves electromagnetic charge, as it must because $SU(2)_L \times U(1)_Y$ breaks to $U(1)_{em}$ after Φ acquires the nonzero vev, implies that the T_{3L} value of field which acquires the nonzero vev will be given by $Q = 0 = T_{3L} + Y/2$. While various entries in this mass matrix will then depend on the isospin and the hyper charge of the Φ,

conservation of the electromagnetic charge will mean that the mass matrix will have a block diagonal form. The same also implies $m^2_{W^1} = M^2_{W^2} = M^2_W$ where M_W is the mass of the $W^\pm$ boson, defined via the last of equations in Eq. 28. The W^3_μ and B_μ will mix. Irrespective of the representation to which the scalar Φ belongs we are interested in the symmetry breaking patterns where $SU(2)_L \times U(1)_Y$ breaks to $U(1)_{em}$ on , on Φ achieving a nonzero vev. Hence one of the eigenvalue of the 2×2 block diagonal matrix aught to be 0. The value of M_Z as well as the ρ parameter will thus depend on the representation of Φ. In fact, it is possible to write a general expression for ρ.

For the present, let us continue with this general form of the matrix without committing to a representation for Φ. Again defining Z_μ, A_μ as in Eq. 29, to be the eigenstates of the above block diagonal mass matrix, it is easy to see

$$\begin{aligned} M^2_\gamma &= M^2_{W^3}\cos^2\theta_W + M^2_B \sin^2\theta_W + 2M^2_{WB}\sin\theta_W\cos\theta_W = 0 \\ M^2_Z &= M^2_{W^3}\cos^2\theta_W + M^2_B \sin^2\theta_W - 2M^2_{WB}\sin\theta_W\cos\theta_W \\ 0 &= (M^2_{W^3} - M^2_B)\sin\theta_W\cos\theta_W + M^2_{BW}(\cos^2\theta_W - \sin^2\theta_W) \end{aligned} \tag{73}$$

This also means $M^2_B + M^2_{W^3} = M^2_Z + M^2_\gamma = M^2_Z$, as it should be since the trace of a matrix is equal to sum of the eigenvalues. Thus we can eliminate M^2_B in favor of M^2_Z. Using Eq. 73, we can easily see that

$$-M^2_{WB} = \frac{M^2_{W^3}(\sin^2\theta_W - \cos^2\theta_W) + M^2_Z\cos^2\theta_W}{2\sin\theta_W\cos\theta_W} = \frac{(2M^2_{W^3} - M^2_Z)\sin\theta_W\cos\theta_W}{\cos^2\theta_W - \sin^2\theta_W} \tag{74}$$

Thus $\cos\theta_W$ can be expressed in terms of $M^2_{W^3}$ and M^2_Z. On comparing Eq. 52 with Eq. 72, we see that for the case of the Higgs doublet we would have

$$\begin{aligned} M^2_{W^1} = M^2_{W^2} = \quad M^2_{W^3} = \tfrac{g_2^2 v^2}{4}, \qquad M^2_B = \tfrac{g_1^2 v^2}{4}, \\ M_{WB} = -\frac{g_1 g_2 v^2}{4} \end{aligned} \tag{75}$$

Using Eq. 74, we then get $M_W = M_Z\cos\theta_W$, precisely the result of Eq. 53. Thus, we see that the $\rho = 1$ prediction is tied to the equality of $M^2_{W^a}, a = 1, 3$ terms in Eq. 72.

In fact a closer inspection of the scalar potential of Eq. 45 reveals that this equality of all $m^2_{W^a}$ is in fact due to an accidental symmetry of the scalar potential for doublet Φ. The doublet Φ contains, in all, four real fields as ϕ^+, ϕ^0 are both complex fields. Writing,

$$\Phi = \begin{pmatrix} \Re e\phi^+ \\ \Im m\phi^+ \\ \Im m\phi^0 \\ \Re e\phi^0 \end{pmatrix} \tag{76}$$

we can see that the scalar potential

$$\begin{aligned} V(\Phi) = -\mu^2\left[(\Re e\phi^+)^2 + (\Im m\phi^+)^2 + (\Re e\phi^0)^2 + (\Im m\phi^0)^2\right] \\ +\lambda\left[(\Re e\phi^+)^2 + (\Im m\phi^+)^2 + (\Re e\phi^0)^2 + (\Im m\phi^0)^2\right]^2 \end{aligned} \tag{77}$$

has an $O(4)$ symmetry under a rotation of the vector Φ of Eq. 76.

Upon SSB, the lowermost component of Φ acquires a non zero vev $\frac{v}{\sqrt{2}}$, whereas all the three components have zero vev.. Hence the scalar potential loses this $O(4)$ symmetry. However, there is still a left over $O(3)$ symmetry corresponding to rotations of the first three components of Φ. among each other. It is this left over $O(3)$ symmetry, called the Custodial Symmetry, which reflects itself in the equality of the masses $M^2_{W^a}$ for $a = 1, 3$ in the matrix Eq. 72, yielding $\rho = 1$.

This also means that even though in the original formulation we had discussed the case of just a single Higgs doublet Φ being involved in the SSB, as long as we use only doublet fields, Eq. 43 is always guaranteed. Of course the statement is true only at the *tree level*. The custodial symmetry, is isomorphic to an $SU(2)$ involving the W^a. This $SU(2)$ is broken by the different masses of the fermions of a $SU(2)_L$ doublet. The value of ρ can change due to contributions coming from loops (as we will discuss in the next section) and also if there exist Higgs belonging to a representation of $SU(2)_L$ other than the doublet.

3.2.7 High energy scattering

Recall the discussion around Eq. 7. We saw there how the postulate of *massive* vector boson was inspired by the demand to restore unitarity to the ν induced processes. For example, the amplitude (say) for $\nu e \to \nu e$ scattering calculated in Fermi theory (current-current interactions) violates tree level unitarity for $\sqrt{s} \lesssim 300 \sim G_\mu^{-1/2}$ GeV. Hence, one could also take this value as an upper bound on the mass of the 'massive' W boson.

However, theories with massive vector bosons have problems with gauge invariance and hence renormalisability. The SSB via Higgs mechanism solved the problem by generating these masses in a gauge invariant manner. This then meant that the theory has renormalisability even with massive gauge bosons. In fact, as we will discuss below, we can see explicitly that gauge invariance also renders nice high energy behaviour to all the scattering amplitudes of the EW theory.

The existence of massive vector gauge bosons restore unitary behavior to processes like (say) $\nu_\mu + e^- \to \mu^- + \nu_e$. But now due to the same non zero mass of the W bosons, amplitudes for processes involving longitudinal W's have a bad high energy behaviour. For example, the matrix element for the process $\nu_e \bar{\nu}_e \to W^+W^-$ through a t-channel exchange of an e, shown in the left panel of Fig. 15, grows too fast with energy and violates unitarity. One can show that

$$\mathcal{M}(\nu_e \bar{\nu}_e \to W^+W^-) \;\sim\; 8\frac{g_2^2}{M_W^2} E p' \sin\theta, \tag{78}$$

where E is the energy of the incoming ν_e and p', θ are the momentum and the angle of scattering of the W boson in the final state. Here we write only the dominant term of the amplitude involving the longitudinal gauge bosons, which is the one with bad high energy behavior. If one does a partial wave analysis of this amplitude, one finds that this amplitude will violate partial wave unitarity, for $s \lesssim \dfrac{M_W^2}{2g_2^2}$ However, what is interesting is that the contribution to the matrix element of the process $\nu_e \bar{\nu}_e \to W^+W^-$, from the s channel exchange of a Z boson, shown in the right panel of Fig. 15 has exactly the same magnitude as the t channel contribution written above but opposite in sign. This happens only if the strength and structure of the couplings of the Z with a ν and W pair is exactly the same as given by the $SU(2)_L \times U(1)$ theory. Thus the violation of unitarity in the amplitude $\nu_e \bar{\nu}_e \to W^+W^-$ due to the longitudinal gauge boson scattering is cured in a gauge theory.

In fact, the GSW model contains more such amplitudes which, in principle, could have had bad high energy behaviour but which are rendered safe by the particle content and the coupling structure of the SM. It was demonstrated [19] that in the GSW model where the masses are generated through SSB by a Higgs doublet (SM), *ALL* such amplitudes satisfy tree level unitarity. In fact the leading divergence of the $\mathcal{M}(WW \to WW)$ which goes like s^2 and hence is much worse, is also cured by the Z exchange contribution and the contribution of the quartic coupling among the W bosons which arise from the non abelian gauge invariance of the theory. Further, the divergent term proportional to s is cancelled by the contribution of the process $W^+W^- \to h \to W^+W^-$, where the Higgs boson is exchanged in the s-channel. Also if one were to calculate high energy behavior of the amplitude $e^+e^- \to W^+W^-$ obtained by replacing the $\nu_e, \bar{\nu}_e$ in the initial state in Fig. 15 by e^-, e^+, then the same cancellation between the divergent parts of the t-channel and s-channel amplitudes is seen to take place.

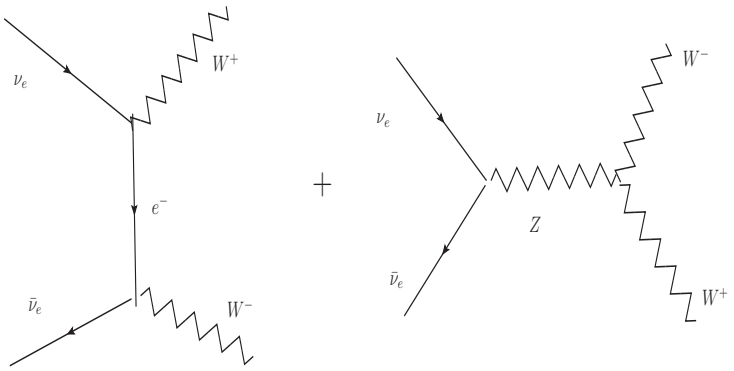


Fig. 15: Gauge theory restoration of tree level unitarity to the $\nu_e\bar{\nu}_e \to W^+W^-$ process.

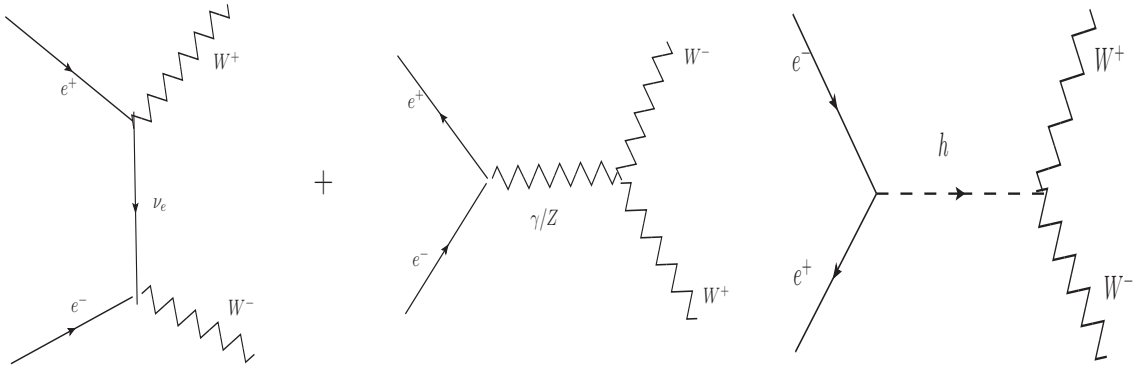


Fig. 16: $e^+e^- \to W^+W^-$ process in the SM

After this observation, a variety of authors [20] investigated the conditions necessary for cancellation of these divergences so that the amplitudes will satisfy tree level unitarity. In fact their analysis indicated that this requires existence of partial wave contributions in the spin 1 and spin 0 channel, with the couplings of these particles exchanged in the s-channel to be precisely those that are given the SM. Recall here that this proportionality of the coupling of the Higgs to the masses of the particles to which it couples is the key prediction of the SSB by Higgs mechanism. The other couplings are of course given by the gauge invariance itself. Thus one could have derived the existence of the Higgs boson as well as the structure of the couplings of the fermions and the gauge bosons to it, without making any reference to the Higgs mechanism and hence the renormalisability.

The fact that the two different requirements, unitarity and renormalisability, lead us to the same result, indicates that there must be a deep connection between the two. In fact, for the $\nu_e\bar{\nu}_e \to W^+W^-$ scattering, there is a residual logarithmic violation of unitarity that is left after all the cancellations, which gets cancelled by the scale dependence of g_2 which is a loop effect which can be computed reliably only in a renormalisable theory.

3.3 Predictions of GSW model

Here we summarize some of the qualitative and quantitative implications of the $SU(2)_L \times U(1)_Y$ invariance. Note that almost all of them are result of the invariance and hence not specific to the actual mechanism of symmetry breaking as long as it preserves the symmetry.

1. First and foremost, this is a unification of weak and electromagnetic interaction: i.e., e, g_1, g_2 all

are of similar order and the apparent difference in strengths of electromagnetic interactions ($\alpha_{\rm em}$ and G_μ), is only caused by the large value of the masses of the weak gauge bosons compared to that of the massless photon. The model predicts existence of a new weak gauge boson Z and that of the weak neutral current (cf. Eq. 38) mediated by it, analogous to the weak charged current of Eq.28 mediated by the W. Further, the strength of this new weak interaction is similar to that of the charged current weak interaction. This is particularly transparent once we use the $\rho = 1$ prediction of the GSW model wherein W/Z masses are generated by SSB using a Higgs doublet.

2. Further, FCNC currents are absent at tree level if and only if all the quarks of a given electrical charge belong to the same representation of $SU(2)_L$. Thus the experimentally observed absence of FCNC implied existence of the charm quark c, in addition to the already known u, d and s quarks. Not only this, one could also 'predict' the mass of the c quark from the measured K_0–$\bar{K}_0$ mass difference.
3. Since G_μ and the electron charge e are measured experimentally, Eq. 41 implies that the model has two free parameters, $\sin\theta_W$ and M_W. If $g_2 = e$, i.e., $\sin\theta_W = 1$, then we get $M_W \sim \mathcal{O}(100)$ GeV. However, when the gauge boson masses are generated through the SSB, M_W can be expressed in terms of G_μ, α_{em} and $\sin^2\theta_W$.
4. The model predicts precise nature of the WWZ coupling, the strength being given by g_2.
5. As Table 5 shows, couplings of all the fermions with the new gauge boson Z, are then determined in terms of $\sin\theta_W$ once the representations of the two gauge groups to which the fermions belong are specified.
6. Requirement of anomaly cancellation, necessary for the renormalisability, predicts that the number of lepton and quark generation seen in nature should be equal. So while the model can not predict how many families of quarks and leptons there should be, it predicts their equality.
7. The conditions of anomaly cancellation and observed closeness of ρ to unity, then gives strong constraints on new particles that one can be added to the spectrum of the GSW model.
8. As already stated above, generation of gauge bosons masses via SSB provides some more relations among physical quantities and hence reduces the number of free parameters of the model to one, that parameter being $\sin\theta_W$.

Thus this model could be easily subjected to experimental tests. This is what we will discuss in the next sections.

4 Validation and precision testing of the SM.

4.1 Early validation.

Historically, the earliest validation of the correctness of the description of the electromagnetic weak interaction in terms of the EW theory, came from the points 1 and 2 in the list given at the end of the last section. By 1972, the renormalisability of the GSW model was proved explicitly [21] and the discovery of weak neutral currents had become very urgent. As we have already noted from Table 5, the NC couplings are entirely decided by the (anti)fermion charge and $\sin\theta_W$. Neutrino scattering with nuclei offer possibility of studying neutral current interactions of quarks. These typically have higher event rates compared to the pure leptonic scattering processes due to the possibility of using nuclear targets. However, analysis of these processes requires an understanding and knowledge of the proton structure. Hence the cleanest probe of the neutral current couplings can come from analysing pure leptonic reactions. We will discuss both of these below.

4.1.1 Discovery of the Weak Neutral Current.

To study the properties of the weak neutral current it was necessary first to establish its existence. To that end, it was necessary to predict the characteristics of the events that would result from interactions

of $\nu_\mu, \bar{\nu}_\mu$ and $\bar{\nu}_e$ beams with electrons, as that would be the cleanest probe. Let us list different types of elastic scattering processes involving just leptons that can take place through weak charged current and neutral current interactions using the ν beams and the electron targets. These are

1. $\nu_\mu + e^- \to \nu_e + \mu^-$, which can take place only through the CC interaction
2. $\nu_\mu + e^- \to \nu_\mu + e^-$ and $\bar{\nu}_\mu + e^- \to \bar{\nu}_\mu + e^-$, which can take place only through the NC interaction,
3. $\nu_e + e^- \to \nu_e + e^-$; $\quad \bar{\nu}_e + e^- \to \bar{\nu}_e e^-$, which can take place both through the NC and CC interactions.

Calculation of the scattering amplitudes of various NC and CC processes listed above (which are depicted in Fig. 8, with appropriate assignments for $f_i, i = 1, 4$) proceeds using the usual rules of field theory. For the low energies of ν- beams that were available then, the $M_W, M_Z \to \infty$ approximation could be used. In situations where both the weak currents (charged and neutral) contribute to a process, the derivation of the effective four fermion interaction in the above limit is a little more involved than our derivations of Eq. 42, but finally leads to very compact expressions very similar to Eq. 42. For the $e^-\nu_e$ scattering mentioned above, for example, the expression resulting from the manipulations is the same as obtained by replacing g_A^e, g_V^e in Eq. 42 by $g_V^e + 1, g_A^e + 1$. Here, we have used $\rho = 1$ prediction of the SM.

Table 7 shows the differential cross-section in terms of the variable $y = \dfrac{E_e}{E_\nu}$ and the integrated cross-section. A few comments are in order. The above expressions use $\rho = 1$ as well as the fact that

Process	$d\sigma/dy$	σ
$\nu_\mu + e^- \to \mu^- + \nu_e$	$A\ s(g_L^\nu)^2(g_L^e)^2$	$A\ s\ (g_L^\nu)^2(g_L^e)^2$
$\nu_\mu + e^- \to \nu_\mu + e^-$	$A\ s(g_L^\nu)^2\left[(g_L^e)^2 + (1-y)^2(g_R^e)^2\right]$	$A\ s\ (g_L^\nu)^2[(g_L^e)^2 + \frac{1}{3}(g_R^e)^2]$
$\bar{\nu}_\mu + e^- \to \bar{\nu}_\mu + e^-$	$A\ s\ (g_L^\nu)^2\left[(g_R^e)^2 + (1-y)^2(g_L^e)^2\right]$	$A\ s\ (g_L^\nu)^2\left[\frac{1}{3}(g_L^e)^2 + (g_R^e)^2\right]$
$\nu_e + e^- \to \nu_e + e^-$	$A\ s\ (g_L^\nu)^2\left[(g_L^e+1)^2 + (1-y)^2(g_R^e)^2\right]$	$A\ s\ (g_L^\nu)^2\left[\frac{1}{3}(g_R^e)^2 + (g_L^e+1)^2\right]$
$\bar{\nu}_e + e^- \to \bar{\nu}_e + e^-$	$A\ s\ (g_L^\nu)^2\left[(g_R^e)^2 + (1-y)^2(g_L^e+1)^2\right]$	$A\ s\ (g_L^\nu)^2\left[\frac{1}{3}(g_L^e+1)^2 + (g_R^e)^2\right]$

Table 7: The differential and total cross-sections for a few $\nu, \bar{\nu}$ induced CC and NC processes, with $A = 4G_\mu^2/\pi$.

values of g_L, g_R for the μ and the e are the same. All the neutrino induced cross-sections are indeed proportional to the square of the centre of mass (com) energy s as we have noted before. The variable y is related to the scattering angle θ in the com frame. One can see after some manipulations that the angular distribution of the scattered charged lepton is different for the case of ν and $\bar{\nu}$. In the first row we have written the cross-section for the CC process $\nu_\mu + e^- \to \mu^- + \nu_e$, so that one can indeed see that the size of the expected cross-sections for the NC processes are of the same order of magnitude as the CC process and depend on $\sin\theta_W$. Note the last two rows of Table 7. As one changes from the $\nu_\mu, \bar{\nu}_\mu$ beams to $\nu_e, \bar{\nu}_e$ beams the factors of $(g_L^e)^2$ in the total cross-section expressions gets changed to

$(g_L^e + 1)^2$. Further, note also the different weights of the $(g_L^e)^2$ and $(g_R^e)^2$ contributions as one changes from ν to $\bar{\nu}$ beams. Both these observations tell us that the contours of constant cross-section for these four processes are ellipses in the g_A–g_V plane with different centers and with major axes of differing orientations. Thus a measurement of these cross-sections can then help us determine g_V^e, g_A^e, albeit upto sign ambiguities.

Note also from the table that as one changes from ν to $\bar{\nu}$, the terms in the angular distribution proportional to $(g_L^e)^2$ and $(g_R^e)^2$ get interchanged. This behavior can be understood very easily in terms of the chirality conservation of the gauge interaction and the angular momentum conservation. As a result, one can write the weak NC cross-sections for all the different pairs of fermions rather easily by inspection. In particular, the same table can be used to calculate the cross-section for the weak NC induced processes with nucleon (nuclear) targets as well. The hadronic weak neutral current events arise from the scattering of the u, d, s quarks in the nucleon (nucleus). In the parton model the net rate is then given by the incoherent sum over all the quarks contained in the nucleon (nucelus). Using the information on the momentum distributions of quarks/antiquarks in the nucleon (nucleus), it is also possible to estimate the expected cross-section. Again these too depend only on $\sin^2\theta_W$ as far as the EW model parameters are concerned.

At the time of the discovery of weak neutral currents in hadronic and leptonic production, theoretical estimates were available for the upper limit on the ratio of neutral current to charged current elastic scattering. This was obtained by using experimental knowledge of the form factor of the proton and neutron. The same was also available for the inelastic process of the inclusive production of hadrons using the language of structure functions of the target nucleus. Two points are worth noting here. While the use of nuclear targets increased the expected rates for NC induced hadron production, establishing that the events are indeed due to weak NC was difficult because of the large neutron induced background. The pure leptonic processes on the other hand, were predicted to be very rare and hence difficult to observe, but could unambiguously prove existence of weak NC as soon as even one event was observed.

Neutral currents were discovered in 1973 in the study of elastic scattering of ν_μ and $\bar{\nu}_\mu$ off nuclear targets [22, 23]. The experiment discovered evidence for the neutral current induced hadronic processes

$$\nu_\mu + N \to \nu_\mu + \text{hadrons}; \;\; \bar{\nu}_\mu + N \to \bar{\nu}_\mu + \text{hadrons}.$$

as well as pure leptonic processes,

$$\bar{\nu}_\mu + e^- \to \bar{\nu}_\mu + e^-,$$

using the giant bubble chamber Gargamelle. In fact the discovery came in an experiment which had been designed to study the charged current interactions:

$$\nu_\mu + N \to \mu^- + \text{hadrons}; \;\; \bar{\nu}_\mu + N \to \mu^+ + \text{hadrons}.$$

Thus the experiment could easily extract the ratio of the CC to NC events, after the observation of NC in hadronic events. The experiment had seen $\mathcal{O} \sim 100$ events of different categories (NC and CC) containing hadrons, with

$$\left.\frac{NC}{CC}\right|_\nu = 0.21 \pm 0.03; \qquad \left.\frac{NC}{CC}\right|_{\bar{\nu}} = 0.45 \pm 0.09$$

As already mentioned, the same experiment also found evidence for the pure leptonic process, where the ν_μ was scattered off the atomic electron. Figure 17, taken from Ref. [23], shows the image of the first unambiguous, weak neutral current event ever observed. The incoming antineutrino, interacts with an atomic electron and knocks it forward. The electron is identified from the characteristic shower created by the electron-positron pairs. This was a considered to be clear evidence for the weak neutral current. The theoretical predictions summarised in Table 7 were used to justify the interpretation. With just one $\bar{\nu}$

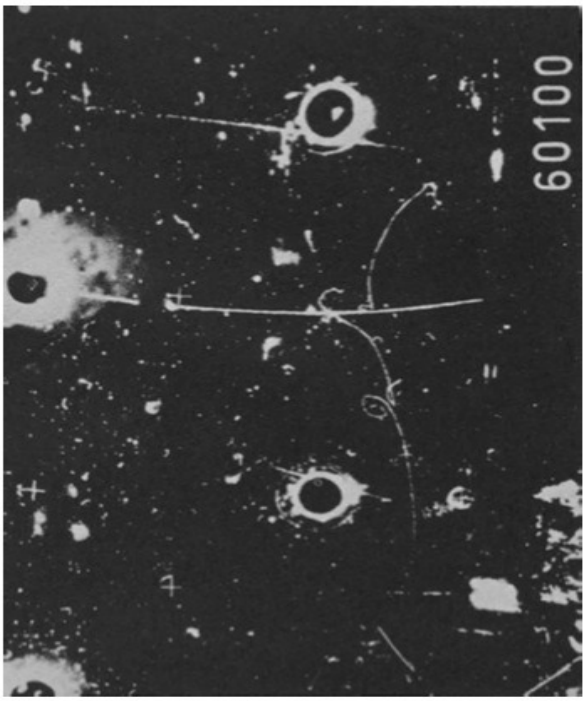


Fig. 1. Possible event of the type $\bar{\nu}_\mu + e^- \to \bar{\nu}_\mu + e^-$.

Fig. 17: Observation of the first leptonic interaction induced by weak neutral current. The incoming $\bar{\nu}$ knocks off the e^-, which then appears as a track accompanied by the shower of e^+e^- pairs the passage of e^- creates. Taken from [23]

event, the experiment could only quote a range $0.1 < \sin^2\theta_W < 0.6$ at 90% c.l. The number of hadronic NC events on the other hand, was big enough to extract a value of $\sin^2\theta_W$ to be in the range of 0.3–0.4. This was the first qualitative validation of the prediction of neutral currents.

4.1.2 Observation of charm with 'predicted' mass

Soon after the observation of the weak neutral current, the charm quark was also discovered with mass very close to that predicted by the analysis of the $\Delta S = 2$, K_0–$\bar{K}_0$ mixing caused by FCNC. We have discussed already details of this prediction in the earlier section. As we understand now, in view of the very large mass of the top quark, it was somewhat 'fortuitous' that the charm quark contribution to the $\Delta S = 2$ mass difference was the dominant one. Be as it may, this was an extremely important second validation of the correctness of the gauge theory of EW interactions based on the gauge group $SU(2)_L \times U(1)_Y$. Note that one of the validation came from tree level couplings and the other from loop induced effects.

4.1.3 Determination of $\sin^2\theta_W$ and prediction of M_W, M_Z.

The same leptonic couplings which contribute to the neutral current scattering processes involving ν's can also make their presence felt in processes like

$$e^+ + e^- \to \mu^+ + \mu^-. \tag{79}$$

This proceeds through a γ^* exchange in the s-channel and a Z/Z^* exchange shown in Fig. 18. Whether the Z will be on shell or off shell of course depends on the com energy. The cross-section for this

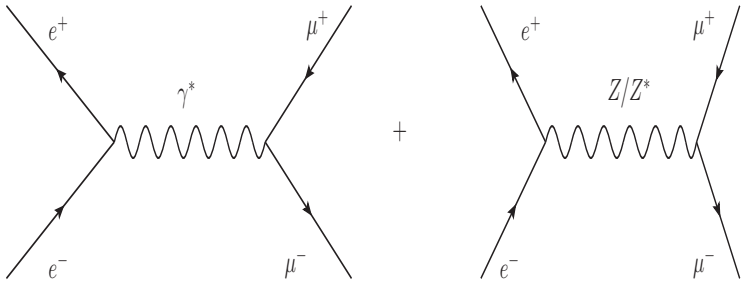


Fig. 18: Weak neutral current effects in $e^+e^- \to \mu^+\mu^-$.

process can be easily computed. Electromagnetic interactions being the same for the left and right chiral fermions, γ^* exchange diagram gives a forward-backward symmetric contribution whereas both, the square of the amplitude of the Z exchange diagram itself and the interference term, will give contributions which are forward backward asymmetric. Hence the presence of the weak neutral current will manifest itself in the form of a forward-backward asymmetry in (say) μ production. Both the size and sign of this asymmetry depends on the centre of mass energy of the process $\sqrt{s} = 2E_b$ where E_b is the beam energy, relative to the mass of the Z boson.

In fact if θ is the angle made by the outgoing lepton with the incoming lepton, then one can show that

$$\frac{d\sigma(e^+e^- \to \mu^+\mu^-)}{d\cos\theta} = \frac{\pi\alpha_{\rm em}^2}{2s}\left[A(1+\cos^2\theta) + B\cos\theta\right] \tag{80}$$

where

$$\begin{aligned} A &= 1 + 2\Re e(\chi)g_V^2 + |\chi|^2(g_V^2+g_A^2)^2;\ \ B = 4\Re e(\chi)g_A^2 \ + \ 8|\chi|^2 g_V^2 g_A^2, \\ \chi &= \left(\frac{G_\mu M_Z^2}{2\sqrt{2}\pi\alpha}\right)\frac{s}{s - M_Z^2 + iM_Z\Gamma_Z}. \end{aligned} \tag{81}$$

Here g_V, g_A denote the (common) vector and the axial vector NC couplings for the e and the μ, Γ_Z is the width of the Z. In the chosen normalisation, deviation of A from 1 and that of B from zero is then indication of the contribution of the weak NC to the process. Both A and B contain terms linear in $\Re e|\chi|$ and g_V^2 or g_A^2. Hence, even for small values of $|\chi|$, both the total cross-section and the angular distribution can be used to probe the weak NC contribution. B is zero without the Z contribution. It is however nonzero for both, the interference terms containing $\Re e(\chi)$ and the square of the Z-exchange diagram alone, containing $|\chi|^2$. Hence the angular distribution contains an asymmetric term at all s. If we analyze these expressions we find that the results for this asymmetry are very different for $\sqrt{s} \ll M_Z$ and $\sqrt{s} = M_Z$.

The forward-backward asymmetry, A_{FB}^μ defined as the ratio of the difference of cross-sections with the μ^- in the forward and the backward hemisphere and the total cross-section, is then expected to be nonzero due to the Z contribution and the interference term. It is clear that this is also the same as charge asymmetry between the muons in the forward hemisphere. Thus one has two asymmetries A_{FB}^μ andA_C^μ:

$$A_{FB}^\mu = \frac{\sigma(\cos\theta^\mu > 0) - \sigma(\cos\theta^\mu < 0)}{\sigma(\cos\theta^\mu > 0) + \sigma(\cos\theta^\mu < 0)}; \qquad A_C^\mu = \frac{\sigma(\mu^-) - \sigma(\mu^+)}{\sigma(\mu^-) + \sigma(\mu^+)}. \tag{82}$$

and these are equal. The reason for the equality of these two asymmetries is the CP invariance of the gauge Lagrangian, even if the Z has parity violating interactions. Using Eqs. 80 and 81 one can calculate the A_{FB}^μ, which in general depends on s. For two different values of s of interest, it can be shown that:

$$A_{FB}^\mu\Big|_{s\ll M_Z^2} = -\frac{3}{\sqrt{2}}\frac{G_\mu s}{e^2}g_A^2\frac{1}{1-\frac{4G_\mu s}{\sqrt{2}e^2}g_V^2}; \qquad A_{FB}^\mu\Big|_{s=M_Z^2} \sim \frac{g_A^2 g_V^2}{\left(g_A^2+g_V^2\right)^2}. \tag{83}$$

In the first case M_Z drops out as we have made an approximation where $s \ll M_Z^2$. In the second case in Eq. 83 , while writing the value for $\sqrt{s} = M_Z$, we have used the fact that $M_Z/\Gamma_Z \gg 1$ and hence the dependency on the precise value of M_Z drops out. The small width is guaranteed by the weak nature of the NC couplings of the Z with the fermions. The factor in the denominator of χ gives a characteristic resonant shape to the cross-section for the process $e^+e^- \to \mu^+\mu^-$, the interference term being negative causing the cross-section to reduce below the value expected for the γ exchange alone and to start rising again as $\sqrt{s}$ approaches M_Z. For $\sqrt{s} \ll M_Z$ the value of A differs from 1, the value expected in QED, by $\left(-\frac{G_\mu s}{\sqrt{2}\pi\alpha_{\rm em}} g_V^2\right)$. Further the coefficient of the asymmetric, linear term in $\cos\theta$ is given by the same expression with the replacement of g_V^2 by g_A^2. Thus it is possible to get information on *both* g_V^2 and g_A^2 from measurements of A and B even with beam energies that are much lower than M_Z. Since $G_\mu \sim 10^{-5}/M_p^2$, the effects can become substantial only when $s \sim \mathcal{O}(10^4\ \text{GeV}^2)$. Indeed the first hints of weak NC in this process were obtained in e^+e^- collisions with $\sqrt{s} \sim 35$ GeV. It is worth noting at this point that the calculation of cross-section for quark (and hence hadron) production via γ/Z exchange proceeds exactly in the same manner, except the expressions will involve g_A^q, g_V^q in addition to g_V^e, g_A^e in Eqs. 80 and 81. All the observations about $e^+e^- \to \gamma/Z \to \mu^+\mu^-$ then apply for the $e^+e^- \to \gamma/Z \to q\bar{q} \to$ hadrons as well.

Note that just like the various cross-sections in Table 7, the asymmetries of Eqs. 82 and 83 too, depend only on one unknown quantity, *viz.*, $\sin^2\theta_W$ through the vector and axial vector NC couplings of the charged lepton. The above expressions tell us therefore, that a study of the leptonic scattering processes given in the Table 7 along with the energy dependence of the FB asymmetry and that of the cross-section for the reaction given in Eqs. 80 and 81, can provide information about $\sin^2\theta_W$ much before reaching the beam energies close to M_Z. If all the measurements of the leptonic cross-sections as well as the asymmetries yielded a unique value of $\sin\theta_W$, which is the only free parameter of the model, this can then provide a quantitative validation of the GSW model. It is interesting to note that the energy dependence of the cross-section $\sigma(e^+e^- \to \mu^+\mu^-)$ can also provide indirect information about M_Z, much before the energy values close to M_Z are reached.

Note that production of hadrons by weak NC processes while being very useful for validation of the weak NC due to the large rates possible with nuclear targets, also needed knowledge about the nuclear structure functions to interpret the data. Both the theoretical and experimental understanding of this structure at that time was somewhat rudimentary. Hence the validation of the SM would be much more unambiguous, if one would extract $\sin^2\theta_W$ using pure leptonic processes alone, *viz.* the ν-charged lepton scattering and e^+e^- collisions.

Fig. 19 shows compilation of such extraction of g_V, g_A and hence $\sin^2\theta_W$ from pure leptonic processes. These results were among the early quantitative validation of the SM. As explained above the leptonic processes were better suited for a clean and unambiguous extraction of $\sin^2\theta_W$. Further, the com energies of the early ν experiments were limited to $s < 200$ GeV^2, whereas the e^+e^- experiments at PETRA at DESY(Hamburg) had $s \lesssim 1400$ GeV^2. The e^+e^- experiments could also probe the NC couplings of the quarks as well, by studying the hadron production along with the $\mu^+\mu^-$ pair production. Thus the information about the weak neutral processes at the e^+e^- colliders was a value addition to the analysis, even though the beam energies were much below than those required to produce an 'on-shell' Z boson. The left panel shows results on the deviation from the QED expectations of the angular distribution for the μ i.e., evidence for both : a nonzero value of B and value of A different from 1. It was indeed comparable to the deviation of few percents to be expected at these energies as was argued above. The plot shows comparisons with predictions of the GSW model (cf. Eq. 83) for different values $\sin^2\theta_W$ showing clear sensitivity to the same. Indeed this as well as measurements of μ charge asymmetry defined in the Eq. 82 for a limited region in the forward hemisphere and the cross-section measurement were used to delineate a region in the g_A–g_V plane that was allowed by the data at 95% c.l. This is indicated by the grey shaded region in the right panel of the Fig. 19. Superimposed on this grey

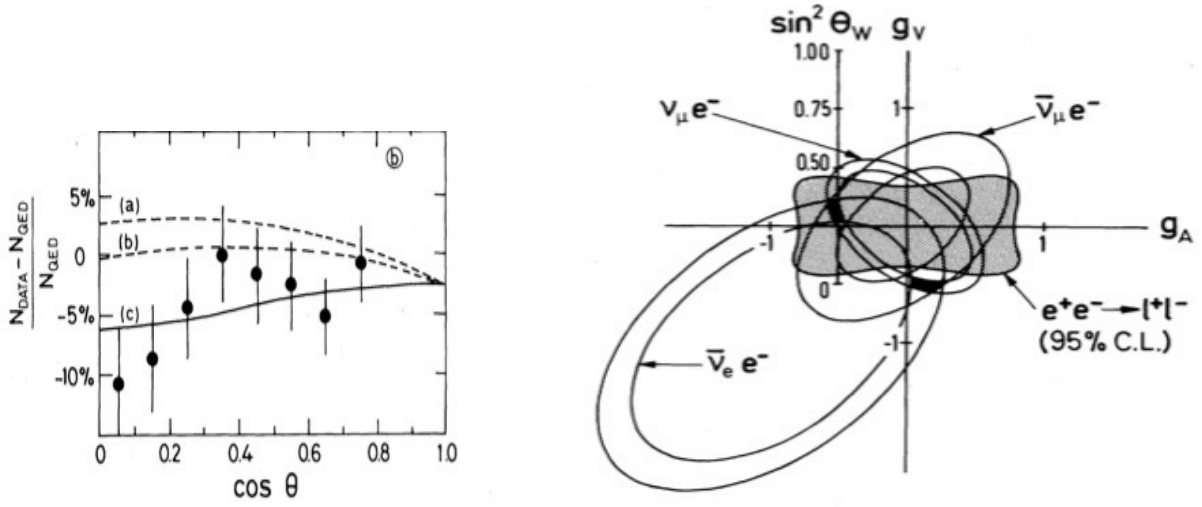


Fig. 19: Quantitative validation of the weak NC. Details of the data taken from [24] are discussed in the text. The left panel shows evidence of asymmetric angular distributions expected from the weak NC contribution. The right panel indicates regions in the g_V, g_A plane and hence values of $\sin^2\theta_W$ extracted from the leptonic data at different level of confidence.

area are also the regions in the same plane allowed by measurements of $\bar{\nu}_e e^-$, $\bar{\nu}_\mu e^-$ and $\nu_\mu e^-$ scattering. We notice from Table 7 that all the cross-section expressions define different ellipses in the g_A–g_V plane. The area between two ellipses is the region allowed at 68% c.l. by the measurement of the cross-section for that particular neutrino scattering reaction.

We see from the right panel that if one uses just the elastic ν-charged lepton scattering data, there is a two fold ambiguity in the values of g_A, g_V that are consistent with the totality of the available data. This is indicated by the two dark black regions. This ambiguity is removed on using the $e^+e^- \to l^+l^-$ data. The solution with negative g_A and positive g_V, corresponding to the dark region in the upper left corner of the grey shaded square region, is chosen uniquely, after we add determination of g_V, g_A from the e^+e^- measurements. This dark region in the upper left corner corresponds to

$$\sin^2\theta_W = 0.234 \pm 0.011. \tag{84}$$

This was the unique value of $\sin^2\theta_W$ consistent with all the 'leptonic' NC measurements mentioned before. One could also use only the e^+e^- data. Combining all the $e^+e^- \to l^+l^-$ measurements with those for $e^+e^- \to q^+q^-$, $\sin^2\theta_W$ was determined to be

$$\sin^2\theta_W = 0.27 \pm 0.08. \tag{85}$$

Clearly the two determinations are consistent with each other. These measurements thus conclusively proved existence of the weak NC as predicted by the GSW model. One could then use the value of $\sin\theta_W$ so determined, to further make predictions for the W, Z masses as well as their phenomenology.

The weak neutral couplings of the electron can also be probed by studying interference between the t-channel γ^* and Z exchange in the Deep Inelastic Scattering (DIS) processes indicated in Fig. 20. This is very similar to the $e^+e^- \to l^+l^-$ case. However, in this case one needs to have longitudinally polarised electron beams, to be able to see the effect experimentally. The diagram with γ^* exchange will give a symmetric result for both left and right polarised e^- but the Z treats them differently. Recall here the different values of g_L^e and g_R^e in Table 5. Thus there will be a polarization asymmetry in the cross-section. At lower energies and hence smaller values of the invariant mass $-Q^2$ of the exchanged γ^*/Z^*, it is the interference term between the two diagrams which dominates the size of the observed

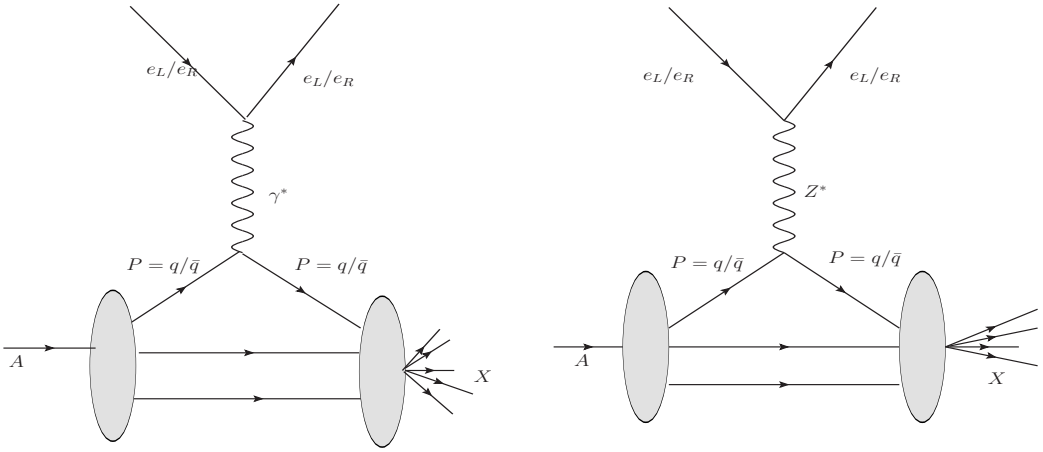


Fig. 20: Weak NC contributions to the deep inelastic scattering with polarised e^- beams.

polarisation asymmetry and hence the evidence for parity violation. The interference effect can be shown to be $\sim G_\mu s$ in this case as well and is linear in g_V^e. As mentioned before, for the value of $\sin\theta_W$ realized in nature the vector coupling of the electron is very small. Hence an asymmetry which is linear in this small parameter, provides a more sensitive probe of g_V^e than the one provided by the asymmetry A_{FB}^μ of Eq. 82. Measurements of this asymmetry also yielded a value of $\sin^2\theta_W$ consistent with the determination from the pure leptonic probes.

Finally the best determination of $\sin^2\theta_W$ came from high statistics data on ν-induced Deep Inelastic Scattering and *polarised* e- Deuterium scattering (both not discussed here at all) and the value was [25]:

$$\sin^2\theta_W = 0.224 \pm 0.015, \rho = 0.0992 \pm 0.017; \quad \sin^2\theta_W = 0.229 \pm 0.009 \ \text{ assuming } \ \rho = 1. \quad (86)$$

In the first case both ρ and $\sin^2\theta_W$ were taken to be unknown and fitted to the data and in the second case ρ was fixed at 1. Thus ρ was determined to be ~ 1 as expected in the GSW model. Assuming this, around 1981 one could then predict using Eq. 55:

$$M_W \simeq 78.15 \pm 1.5 \text{ GeV}; \qquad M_Z \simeq 89 \pm 1.3 \text{ GeV}. \quad (87)$$

This then sets the goal posts to design experiments which could produce W, Z directly and study them. In principle, the predictions above receive radiative corrections. A more accurate prediction would require, for example, discussion of radiative corrections to the couplings involved in the relations given by Eq. 32. We will come to that in the next subsection.

So the take home message of the above discussion is that the early ν experiments as well scattering experiments with polarised electron beams and nuclear targets, along with the $e^+e^- \to l^+l^-$ experiments, tested the structure of the NC couplings of the leptons AND those of the quarks predicted by the GSW model. The experiments conclusively proved that all the measurements were consistent with a *unique* value of the one undetermined parameter of the model $\sin^2\theta_W$. This then also predicted a narrow range of possible masses for both the W and the Z bosons. Inter alia, these measurements also established $\rho \simeq 1$, consistent with the GSW prediction again. Thus at this stage, apart from the direct verification of the tree level ZWW coupling which must exist in this gauge theory, all the other tree level predictions of the model seemed to have been tested.

Given the knowledge of the quark content of the p available from the DIS experiments, it was also possible to predict the rate of production of these bosons in the process

$$p + \bar{p} \to W + X \to l + \nu_l + X; \qquad p + \bar{p} \to Z + X \to l^+ + l^- + X.$$

In fact the CERN super proton synchrotron (SPS) was converted into $Sp\bar{p}S$, to collide protons on antiprotons, so as to have enough energy to produce the W, Z in the $p\bar{p}$ collisions. The observation of the W and the Z bosons in the UA-1 and UA-2 experiments [26, 27], with mass values and production rates which agreed with these predictions, was a very important step in confirming the correctness of the GSW model. Later data confirmed the $V-A$ coupling of the W bosons to fermions from the angular distribution of the events, even though the original observation had only a handful of these: 6 in UA-1 and 4 for UA-2.

The masses of the W and the Z measured in the UA2 experiment [27], for example, were

$$M_W = 80 + 10 - 6 \text{ GeV}; \quad M_Z = 91.9 \pm 1.3 \pm 1.4 \text{ GeV}.$$

The larger errors for M_W reflect the uncertainties in the measurement of 'missing' transverse momentum due to the ν which evades detection. For M_Z, the first number indicates the statistical error and the second systematic. The use of final state containing leptons allowed for much more accurate determination of the invariant mass in the case of the Z boson. These masses were certainly consistent with the predictions: see, for example, Eq. 87. One can in principle extract ρ AND $\sin^2\theta_W$ from this 'direct' measurement of masses (in particular the accurate measurement of M_Z) and compare these with the values obtained from the earlier 'indirect' information from ν scattering, for further tests of the SM. This already used the more accurate predictions using energy dependence of the couplings as well EW corrections to the weak processes used to extract $\sin^2\theta_W$. We will discuss this in the context of precision testing of the SM.

4.2 Direct Evidence for the ZWW coupling.

Before moving on to the discussion of calculation and validation of loop effects in the precision measurements of the EW observables, we need to discuss the validation of the existence of another tree level coupling of the gauge bosons, *viz.*, the triple gauge boson ZW^+W^- coupling which is characteristic of the non abelian nature of the gauge theory. As already discussed, contribution of the Z exchange diagram is crucial in curing the bad high energy behavior of the $e^+ + e^- \to W^+W^-$ cross-section. W^+W^- pair production in e^+e^- collisions was studied at LEP-II where the centre of mass energy was increased from the Z-pole value of 91 GeV to the two W threshold of 161 GeV and then finally to 209 GeV. Fig. 21 shows the LEP-II data along with the theory prediction. The data is well described by the solid line which represents the sum of the contribution of the ν_e exchange diagram and Z/γ exchange diagrams shown in the left and the central panel of Fig. 16. One sees that the contribution to the cross-section of just the ν_e exchange diagram of the left most panel, shown by the blue dashed curve, rises very fast with energy. The cross-section after including contribution of the s-channel γ exchange alone, where the ZWW coupling is put to zero in the diagram in the central panel of Fig. 16, is shown by the red dashed curve. This addition tames the bad high energy behavior to some extend but not completely. Only after adding the s-channel Z-exchange diagram does the cross-section have a good high energy behavior, shown by the blue-green solid curve which also describes the data well. Thus we see that the temperate energy dependence of the $e^+ + e^- \to W^+W^-$ cross-section shown by the data, is 'direct' proof of the ZW^+W^- triple gauge boson coupling.

The threshold rise of this cross-section also offers an accurate determination of W mass and the width [28]:

$$M_W = 80.376 \pm 0.033 \text{ GeV}, \Gamma_W = 2.195 \pm 0.083 GeV.$$

The same experiment offered a precision measurement of the hadronic decay width of the W as well. These measurements served later as an input to the precision analysis of the EW observables which we will discuss in the next section.

Note further also that since the energy dependence of the total cross-section is crucially decided by the ZW^+W^- coupling, it is possible to use the energy dependence and the angular dependence of

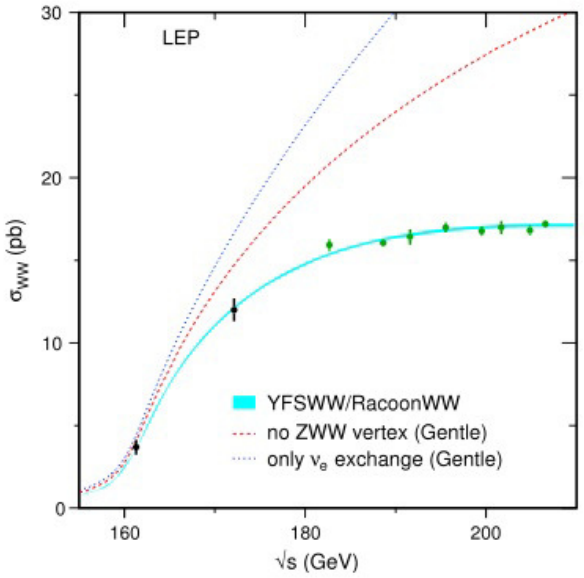


Fig. 21: Energy dependence of the W^+W^- cross-section at LEP-II. Taken from [28].

the process to probe any possible deviations of the ZWW vertex from the SM structure and value. This process can therefore be successfully used to look for deviations of this coupling from the SM prediction. In view of the important role played by the ZWW coupling in curing the bad high energy behavior of the W-pair production cross-section, it is theoretically very important to probe its possible deviations from the SM predictions so as to get indications, if any, of the physics beyond the SM (BSM physics). Measurements of the cross-section and angular distributions of the produced W at LEP-II, constrained strongly any anomalous ZWW couplings; i.e., couplings which differ from the SM in either structure or strength.

4.3 Precision testing of the SM

Thus we see that the various lepton-lepton and lepton-hadron scattering experiments along with the $p\bar{p}$ experiments helped establish the correctness of GSW model predictions at the tree level. These tested the tree level SM predictions for the new NC couplings of the Z boson with all the known fermions as in terms of the single 'free' parameter of the model. The prediction of $SU(2)_L$ symmetry for the structure and strength of the ZWW vertex was also tested. Last but not the least the experiments also tested the correctness of the tree level predictions for the W and Z masses. This indeed established the $SU(2)_L \times U(1)_Y$ structure of the EW gauge theory. However, even with the somewhat imprecise determined values of W, Z masses, the need for including the effects of loop corrections, an essential feature of QFT's, on all these tree level predictions was already clear. Since the effect of radiative corrections on the extraction of $\sin^2\theta_W$ is different for different processes, it is necessary to correct the experimentally extracted value for these effects, before the $\sin^2\theta_W$ extracted from various observables can be compared at high precision.

4.3.1 Radiative corrections and ρ/$\sin^2\theta_W$ determination.

In case of the SM, a QFT with SSB, renormalisability of the theory guarantees that the loop corrections to the tree level relations such as given by Eqs. 32,43 and 53, will be finite and can be computed order by order in perturbation theory. Precision measurements can then test these corrected relations and hence the correctness of these calculations of loop effects. This can then help establish the renormalisability of

the $SU(2)_L \times U(1)_Y$ gauge theory of EW interactions. Below follows an extremely sketchy discussions of the issues involved.

Some of the one loop diagrams contributing to the corrections to the vertices and two point func-

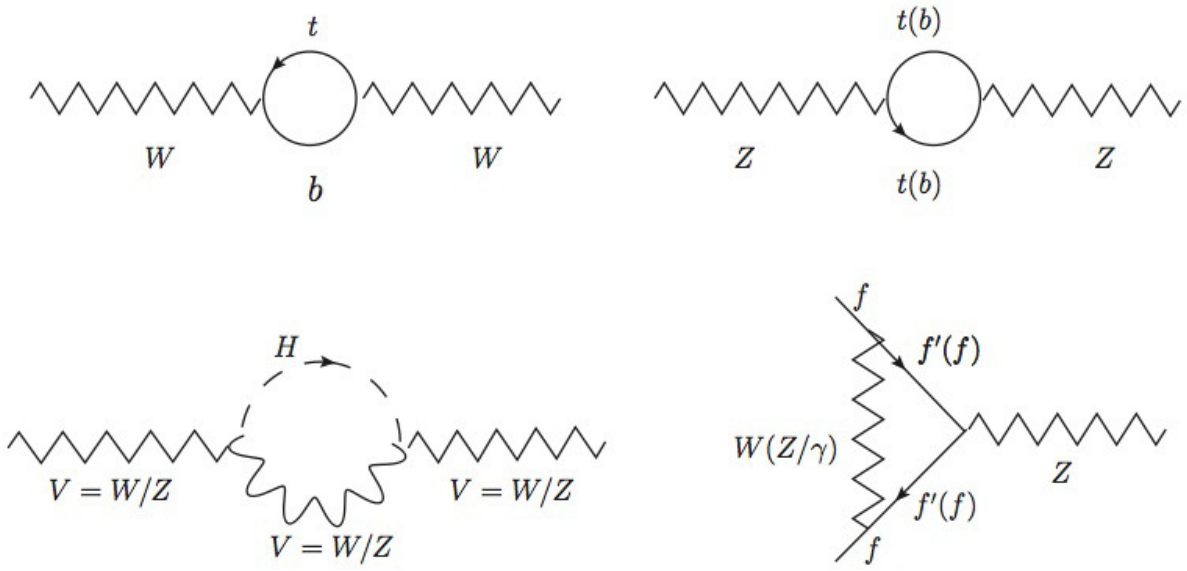


Fig. 22: Some of the one loop EW corrections to vertices and two point functions in the SM.

tions are shown in Fig. 22. The two diagrams in the top row and the diagram on the left in the lower panel are the ones that need to be considered while calculating the loop corrections to the masses M_W, M_Z. The diagram on the right in the lower panel is an example of diagrams that give rise to corrections to the $Zf\bar{f}$ vertex. The dominant corrections come from loops containing quarks of the third generation viz. t, b. We already notice that corrections to the W and the Z mass will be different, since the former involves a tb loop where as the latter involves the $t\bar{t}, b\bar{b}$ loops. As a result the corrections to $\sin^2\theta_W$ from these diagrams, for example, will be different for the CC and NC processes. Let us recall Eq. 55. We have used Eqs. 32 and 53 in deriving this. One needs to take into account radiative corrections to the weak processes used to extract $\sin^2\theta_W$ as well as the energy dependence of the couplings and hence of $\sin^2\theta_W$ obtained via Eq. 32. The latter too is an integral part of QFT. The extraction of $\sin^2\theta_W$ from weak processes, taking into account all the weak corrections yielded [29]

$$\sin^2\theta_W(M_W) = 0.215 \pm 0.010 \pm 0.004.$$

In Eq. 55 one now needs to use $\alpha_{\rm em}(M_W) = 1/127.49$ instead of the value $\alpha_{\rm em} = 1/137.03$ used therein. The expression for $M_W(M_W)$ then becomes

$$M_W(M_W) = \sqrt{\frac{\pi}{\sqrt{2}G_\mu}\frac{\alpha_{\rm em}(M_W)}{\sin^2\theta_W(M_W)}} = \frac{38.6}{\sin\theta_W(M_W)}\ \text{GeV}. \tag{88}$$

This then gives,

$$M_W = 83.5 \pm 2.2\ \text{GeV};\ \ M_Z = 94.2 \pm 1.8\ \text{GeV}. \tag{89}$$

Thus loop effects change the predicted values from those in Eq. 87 by $\mathcal{O}(\sim 5\%)$. This sets the scale for the precision with which one needs to measure the values of the masses of the W, Z to be able to test theory at loop level. The UA-1 and UA-2 measurements were clearly consistent with these predictions

within the accuracy of the measurement as well as predictions. In Ref. [27] these loop corrected predictions for M_W, M_Z were used to extract both $\sin\theta_W(M_W)$ and ρ just from the measured masses of the W, Z in the UA-2 experiment, yielding

$$\sin^2\theta_W = 0.226 \pm 0.014, \quad \rho = 1.004 \pm 0.052.$$

This value of ρ is consistent with the expectation of the SM i.e., the GSW model where W/Z masses are generated via SSB. These values are also consistent with the corresponding determinations from the lower energy ν experiments (cf. Eq. 86). Agreement of these two independent determinations of ρ and $\sin^2\theta_W(M_W)$ from two completely different sets of measurements, already showed consistency of the measurements with theory predictions at loop level.

Diagrams shown in Fig. 22 cause ρ to change from 1, the prediction at tree level, since the corrections are different for M_W^2 and M_Z^2. In fact, one can write

$$\Delta\rho = \frac{\Sigma_Z(0)}{M_Z^2} - \frac{\Sigma_W(0)}{M_W^2},$$

where Σ_V, $(V = W/Z)$ are the one loop corrections to the propagator. As emphasized above these are different for the W and the Z and hence $\Delta\rho$ is different from 0. At one loop one gets, keeping only the dominant corrections $\propto M_t^2$,

$$\rho_{\rm corr} = 1 + \Delta\rho \simeq 1 + \frac{3G_\mu M_t^2}{8\pi^2\sqrt{2}} \tag{90}$$

Thus one sees that the relation $\rho = \dfrac{M_W^2}{M_Z^2\cos^2\theta_W} = 1$ gets corrected by loop effects. The corrections are finite as advertised before: a result of the renormalisability of the EW theory. Assuming the (at that time) unknown M_t to be as large as the largest mass in the theory, $\sim \mathcal{O}(M_W)$, one finds corrections to the tree level value of unity of ρ, to $\sim$ few parts in 1000. Thus one would need a high precision measurements of M_W, M_Z to get a precision value of ρ which can then be contrasted with above prediction given in Eq. 90. This can then be used to estimate M_t and comparing it with the experimentally observed value of the t quark mass would then constitute a precision test of the SM.

In reality, indeed this is what happened. Recall the discussion around Fig. **??**. The precision measurements at the Z pole in $e^+e^- \to Z \to f\bar{f}$, to be discussed momentarily, indicated a value for the top mass $M_t \simeq 2M_W$ *before* the top quark was actually discovered. Agreement of the measured mass of the t at the Tevatron with this value was then a big success story, testing the SM at loop level. For the much higher value of the mass that the t quark has in real life compared to the M_W taken in the numerical estimation above, corrections to ρ in reality are about 1 part in 100 and hence measurable in precision experiments. For future reference, let me also add here that the corrections to M_V^2, from the third diagram in Fig. 22 involving the VH loop, depend on the Higgs mass M_h only logarithmically.

A detailed discussion of the theoretical significance of the all important quadratic dependence of these corrections on M_t, the logarithmic dependence on M_h and the non decoupling nature of the corrections to the $Zb\bar{b}$ vertex from the $t\bar{t}$ loop, are beyond the scope of the discussion in these lectures. The former comes from violation of the $SU(2)_L$ invariance, reflected in the mass difference between the two members of the doublet : the t and the b. $\Delta\rho$ is in fact proportional to $M_t^2 - M_b^2$. The loops involving h and the V give contributions to $\Delta\rho$ which depend on the Higgs mass, but the accidental Custodial Symmmetry (cf. section 3.2.6), guarantees that this dependence will be only logarithmic. This is consistent with the so called Veltman screening theorem [30]. The corrections to the $Zb\bar{b}$ vertex, originating from the triangle diagram, one of which is shown in Fig. 22, also depend on M_t quadratically. This quadratic dependence, on the other hand has a different source. It arises from contributions of the longitudinal W bosons in the loop. In a non-unitary gauge this can be seen as coming from the unphysical Goldstone bosons $\phi^\pm$, which are 'eaten up' to become the longitudinal degree of freedom

of the W-boson. This then clearly explains the non decoupling nature of the correction, coming from the proportionality of $t\phi^{\pm}$ coupling h_t or equivalently M_t. Even when we do not discuss these issues in detail, suffice it be said that the M_t^2 dependence of the vertex correction is the tell tale sign of the SSB via the Higgs mechanism. Since the origins of the M_t^2 dependence, or equivalently the non-decoupling nature of the corrections, are quite different for the $\Delta\rho$ and $\delta g_{Z\mu\mu}$ and further only the $\Delta\rho$ receives contribution from the Higgs, it is quite important to confirm both of these independently. Let us now follow the story of precision measurements and comparison with the precision predictions further.

Note here that these corrections can be calculated only if theory is renormalisable. The renormalisability of a gauge field theory with SSB was proved by 't Hooft [21]. This theory necessarily has a physical scalar, the Higgs boson in the spectrum. As we will see shortly, the precision measurements at the LEP-I of the Z properties along with weak neutral current couplings of all the fermions, as well as precision measurements of the properties of the W at LEP-200, tested these corrections. A test at the loop level of the various relations such as Eq. 32 or Eq. 53, could then indicate the need for a finite mass for the Higgs and thus could be an indirect proof for the Higgs! However, we have seen that even with a quadratic dependence of $\Delta\rho$ on M_t and the large mass M_t, the effects are only 1 part in 100, it is clear that with the logarithmic dependence of these corrections on M_h, this program would require indeed very high precision measurements.

4.3.2 Precision measurements at LEP

Let us first begin by a discussion of precision measurements of the mass and the coupling of the Z boson at LEP 1 and the SLC in $e^+e^- \to Z \to f\bar{f}$. The four LEP experiments studied decays of about 17 Million Z, whereas the SLC studied about 600,000 Z decays, but with polarized e^+/e^- beams. These precision studies of the Z have been summarised in Ref. [31]. At the end of the day these experiments determined the mass and the width of the Z boson and also the values of ρ and effective value of $\sin^2\theta_W$, to a great accuracy using only the leptonic sector. The use of 'effective' implies that radiative corrections have been suitably included while extracting these values.

$$\begin{aligned} M_Z &= 91.18750.0021 GeV, & \Gamma_Z &= 2.49520.0023 GeV, \\ \rho_l &= 1.00500.0010, & sin^2\theta^{eff}_{lept} &= 0.231530.00016. \end{aligned} \tag{91}$$

As already explained these high precision measurements require also high precision calculations, to test the SM at high accuracy. Higher order QCD corrections play a highly important and nontrivial role while using results from the hadronic decays of the Z. One also requires an excellent understanding of QCD to calculate correctly the observables from quark final states in terms of what the detectors actually observe *viz.* the jets. This ushered in an era of extremely close and extensive collaboration between experimentalists and theorists resulting in a number of LEP Yellow Reports. These provide the best summary of both the theoretical and experimental issues involved in studies at LEP.

Fig. 23 shows a compilation of the cross-section for the process $e^+ + e^- \to \text{hadrons}$, spanning the entire energy range from PEP/PETRA to LEP II. Solid line is theory prediction, including the electromagnetic and the QCD radiative corrections. Recall the expression for the cross-section for $e^+e^- \to \mu^+\mu^-$ given in Eq. 81. The initial fall off of the cross-section reflects the $\frac{1}{s}$ dependence of the first γ exchange diagram in Fig. 18. One can then see the onset of the rise in the cross-section due to interference between the γ and Z exchange contributions. Recall that it is these interference terms, at energies quite far away from the Z resonance, that had allowed the first glimpse of effects of weak neutral current in the process $e^+e^- \to \mu^+\mu^-$. Thus we see that the Z resonance makes its presence felt much before the resonant energy is reached, by just the shape of the cross-section curve. This line shape of the Z resonance depends on Γ_Z, M_Z, partial decay width $\Gamma(Z \to f\bar{f})$ and through them on g_V, g_A of the electron and the fermions in the final state being considered. The extremely accurate measurements of M_Z, Γ_Z mentioned above, were extracted by fitting the shape of this curve near resonance, taking

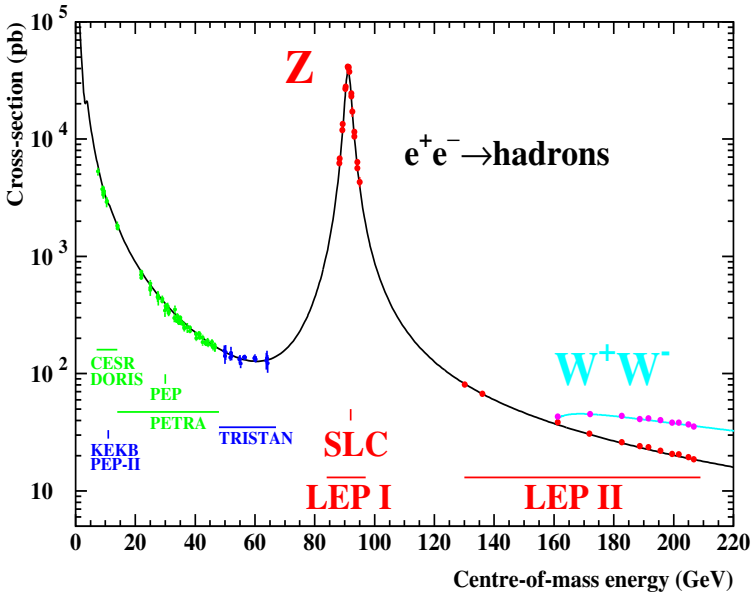


Fig. 23: The figure shows summary of the data on $e^+e^- \to$ hadrons over a wide energy range taken from [31].

into account effects such as the initial state radiation etc. This precision study of the line shape of Z was made possible by the unprecedented energy resolution of the collider LEP-I. The thin solid line is then the theoretical prediction for the cross-section including the QED and QCD radiative correction. The asymmetric shape of the curve near the resonance is the effect of the initial state radiation. The agreement between the data and theory needs no comment.

Recall now the discussion in Sec. 4.1.3 and Eqs. 81 -83. One can extend constructions of these asymmetries of Eqs. 81-83, for all the fermionic final states accessible in the Z decay, *viz.* the leptons e, μ, τ and the quarks b, c. Looking at the expressions in Eqs. 81 – 83 one can see that a precision measurement of these asymmetries as well as partial widths, lead to an accurate determination of g_V^f, g_A^f. The Z-decay data from SLC, which employed linearly polarised e^-/e^+ beams, allowed for constructing polarisation asymmetries just like the forward-backward asymmetry of Eq. 83. This too is a measure of parity violation, with the additional advantage that it involves g_V linearly instead of the quadratic dependence in Eq. 83. This linear dependence is similar to the case of polarization asymmetries in case of polarized electron-Deuterium scattering mentioned before. Recall also that for the value of $\sin^2\theta_W$ of Eq. 86 which is rather close to 0.25, the vector coupling of the electron involving $(4\sin^2\theta_W - 1)$ is very small. Hence this linear dependence of the asymmetries on g_V allowed the experiments at the SLC to reach a competitive accuracy for the extraction of g_A, g_V with the much smaller luminosity and hence smaller number of the Z decays (600000 versus 17 million at LEP) available there.

Fig. 24 shows values of g_V^e, g_A^e obtained using the LEP-I data, juxtaposed with the data from elastic ν scattering from 1987. The latter is a more refined version of the of the plot of Fig. 19. To truly appreciate the phenomenal improvement, compare the size of the region in the g_V, g_A plane selected by all the measurements (shown in an inset at the left of the figure, blown up by roughly a factor of 1000) with the size of the corresponding region in Fig. 19. Thus we see that at the Z pole the weak NC couplings of the Z with the fermions, were tested to about one part in 1000.

It goes without saying that with such precision in measurements, if one were to repeat the earlier exercise of extracting the value $\sin^2\theta_W, \rho$ from them, such as given in Eq. 91, one HAS to use theoretical predictions which include all the relevant higher order corrections. This was already discussed in Sec. 4.3.1. Since these corrections have a dependency on the masses of the particles like the W, t

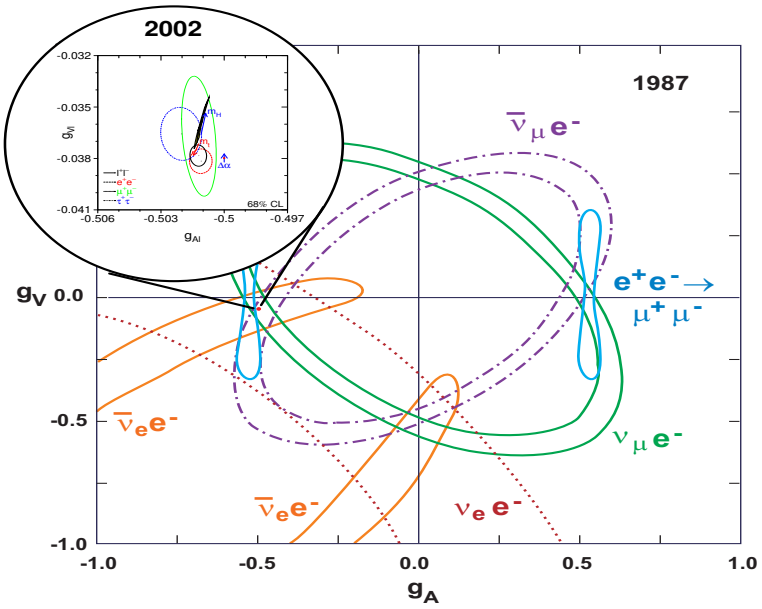


Fig. 24: the plot shows determination of g_V, g_A of the electron using Z decays taken from [31].

and the Higgs, if the measurements are precise enough then they can be sensitive to these masses. We already saw this for the mass of t quark and the radiative corrections to the ρ parameter. The precision measurements of EW observables then indicate 'indirectly', *in the framework of the SM*, the values of the masses of these particles preferred by the precision EW data. A comparison of these masses determined 'indirectly', with the ones measured directly, can then be a powerful precision test of the SM.

4.3.3 Precision testing and indirect bounds

Let us describe the logical steps in such a program to perform precision testing of the SM. In principle the EW part of the SM has following free parameters: g_1, g_2, v and λ. In addition to this of course there is the QCD coupling g_3, the nine masses (or equivalently the Yukawa couplings) of the massive charged leptons and quarks, the four parameters of the CKM matrix and the strong phase θ_{QCD}. At tree level all the couplings of the gauge bosons to fermions as well as to each other and their masses are completely given in terms of the first three parameters in this list, *viz.* g_1, g_2 and v. In section 3.2.2 we already discussed an analysis where we traded these three for the more accurately known $\alpha_{\rm em}$, G_μ and one free parameter $\sin\theta_W$ (cf. Eq. 55). With the very precise knowledge of M_Z provided by the LEP-I, it made sense to trade the g_1, g_2 and v for $M_Z, \alpha_{\rm em}$ and G_μ. As before, one can then use the relationships such as given by Eqs. 32, 53 etc., of course corrected for radiative effects, to express all the EW observables as functions of these three chosen quantities.

A really large number of EW observables have been measured very accurately, beginning from the total width of Z boson, Γ_Z, the various forward-backward and polarisation asymmetries on the Z-pole, masses M_W, M_t, polarised e-Deuterium scattering, atomic parity violation etc. All these observables depend on G_μ, M_Z and $\alpha_{\rm em}$ through their dependencies on g_A^f, g_V^f, M_V as well as on α_s and M_t, M_h through the higher order QCD and EW corrections.

Precision calculation for all these EW observables, including the 1 loop EW radiative corrections in the framework of the SM, are available. The idea is to make then a fit to the measured values of the EW observables and test the SM predictions. In these fits, one keeps M_t, M_W and M_h as free parameters. As already noted the radiative corrections depend on M_t quadratically and M_h logarithmically. Then compare the M_W, M_t values so obtained with experimentally determined values of the same, thus

providing a test of the SM. Afterwords one can perform the exercise by varying the Higgs mass, find the value of M_h that minimises the χ^2 and then find the limits on the Higgs mass for which the data will be consistent with the predictions of the SM.

Fig. 25, taken from the url of the LEP EW working group [32], shows the result of such an exercise. The figure lists the measured values of a variety of EW obsevables, most of which we have discussed. The various R-ratios: R_b, R_c, R_l etc. are a measure of the relative production of the various final states and hence of the partial decay width of the Z into them. $A_l(P_\tau)$ is the polarisation asymmetry for the τ's produced in $e^+e^- \to Z \to \tau^+\tau^-$ on the Z–pole. The second column shows the result of the SM fit for the observable and the third column the pull which is the difference between the measurement and the fit value normalized by the error of the measurement. The pull is less than three for all the observables and above 2 for only one of the measurements *viz.* A^b_{FB}. This particular fit is the last one **before** the

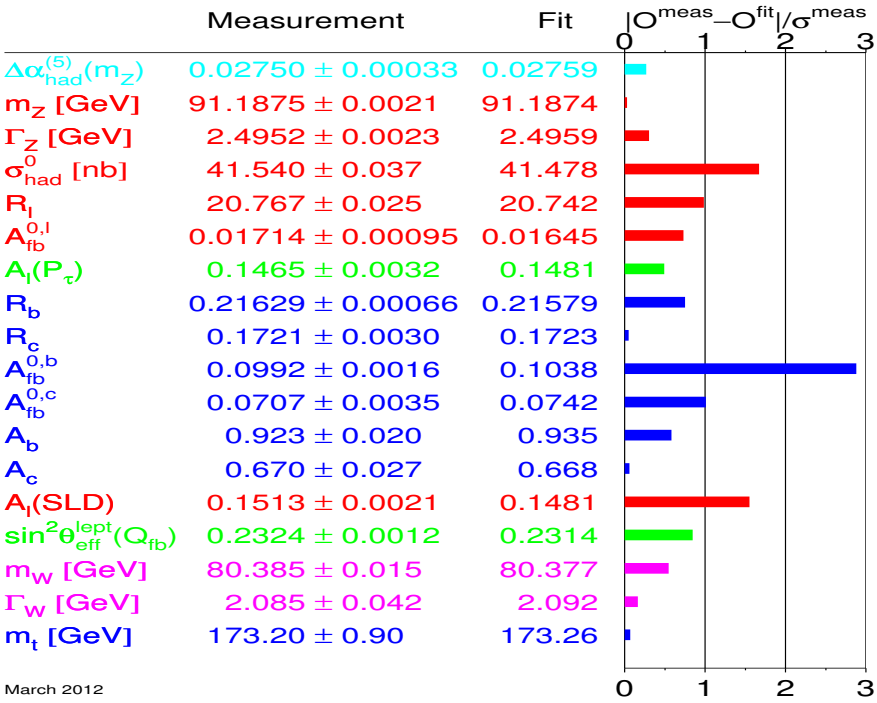


Fig. 25: Pull for the SM fit for the totality of the EW precision observables. Taken from [32].

discovery of the Higgs at the LHC, using the most accurate measurement of M_W from the Tevatron, which has an error of 0.15 GeV, again a 'one per mille' measurement. The χ^2 of this fit is not very small, mainly due to the discrepancy between the best fit values and measured values for A_b from LEP as well as at the SLC. Hence before the 'direct' discovery of the Higgs there were a few physicists who used to be a little uncomfortable about the goodness of the fit and accepting this as 'the proof' for the correctness of the SM at loop level.

Note the values in the last two rows. The measured values and the best fit values of M_W, M_t agree with each other to a great precision and the pull is is rather small, providing thus a stringent test of the SM at loop level. This is the agreement between the M_t predicted 'indirectly' from the LEP EW precision measurements and the 'direct' measurement from the Tevatron, that was alluded to before a few times. In fact this spectacular agreement was the QFD (Quantum Flavour Dynamics) equivalent of testing the $(g-2)_\mu$ prediction with the measurement in QED. The important role played by renormalisabilty and loop corrections in this context can be understood by doing a small numerical exercise of predicting

M_W from the very accurately measured values $\alpha_{em} = 1/137.0359895(61)$, $G_\mu = 1.16637(1) \times 10^{-5}$ GeV^{-2}, $M_Z = 91.1875 \pm 0.0021$ GeV and the tree level relations given by the SM among these quantities and M_W. Notice that Eq. 55 can be written as,

$$\frac{G_\mu}{\sqrt{2}} = \frac{g_2^2}{8M_W^2} = \frac{\pi\alpha_{\rm em}}{2M_W^2(1 - M_W^2/M_Z^2)}$$

by using the tree level relation $M_Z = \dfrac{M_W}{\cos\theta_W}$. This gives, $M_W^{tree} = 80.939$ GeV. Compare this now with the value of M_W given in the second column of Fig. 25, $M_W^{expt} = 80.385 \pm 0.015$ GeV. Of course, this points out the need for calculating loop corrections to the tree level relations. Renormalisability guarantees that all the corrections are finite and can be computed. Hence the value of M_W obtained 'indirectly' from the fits using theoretical predictions which *include* these loop corrections, then famously agrees with the 'direct' measurement as shown in Fig. 25. Agreement with the SM prediction would have been impossible unless the predicted values included higher order corrections calculated in perturbation theory.

The fit values and the pull for M_t, M_W depends on the value of M_h, albeit very weakly, due to the logarithmic dependence on M_h of the EW corrections to M_W, M_Z etc. Some of these effects can

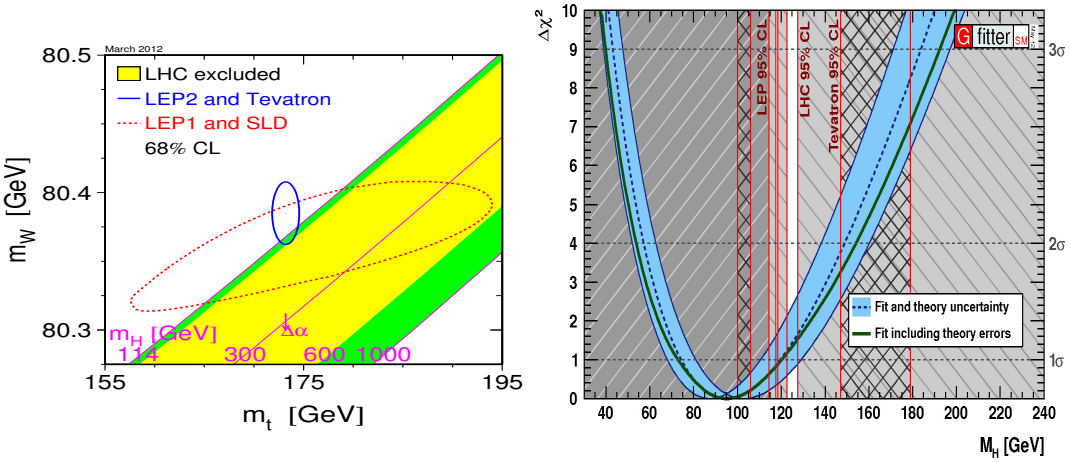


Fig. 26: Left panel shows the dependence on M_h of the M_W–M_t values obtained from the EW precision data. Taken from [32]. The right panel shows the status of the 'indirect' limits on M_h obtained by fits to the EW precision data. This is taken from [14]. Both these are from the eve of the Higgs discovery, March 2012.

be seen from the two panels in Fig. 26. The plot in the left panel shows the dependence of the fit values for M_W, M_t for different values of M_h. The long lopsided ellipse used the EW observables measured at LEP-I and the SLC, to determine allowed regions in the M_t–M_W plane at 95% c.l. Using the M_W measurements at the LEP-II/Tevatron as input, one now obtains the small blue ellipse which is consistent with the precision measurements. The dark green (grey) region and the large red ellipse show that with results from LEP-I alone, the measurements were not sensitive to M_h at all. On the other hand, the highly accurate LEP-II/Teavtron measurements of M_W and the Tevatron measurement of M_t is consistent with somewhat small values of the Higgs mass at the left most boundary of the green(grey) region. This was also consistent with the exclusion (from direct searches at the LHC) of a SM Higgs over a very large range as indicated by the M_h values labeling the inclined lines in the region shaded in yellow (a shade of lighter gray).

The right panel shows the same information in a different format, where we show a plot of $\Delta\chi^2$ as a function of M_h. In fact the fact that this minimum of $\Delta\chi^2$ occurs at a nonzero, finite mass M_h is

already an indication of the 'existence' of the Higgs and hence a feather in the cap of the SM. The dotted and solid black lines are the best fit with and without including the theory errors. The region shaded in light blue (grey) indicates effect of the theoretical uncertainties as well as uncertainties in the EW fit. In the absence of any information from 'direct' searches for the Higgs, the indirect constraints will allow a region around the minimum of χ^2 ($M_h \simeq 90 - 100$ GeV) upto M_h values where $\Delta\chi^2$ is 9: the 3σ value. Remaining values of M_h will be disfavored by this 'indirect' search. The $\Delta\chi^2 \leq 9$ corresponds to an allowed mass range $40 - 45 \lesssim M_h \lesssim 180 - 200$ GeV at 3σ. However a lot of this 'allowed' region is ruled out from direct searches at the LEP, at the Tevatron and at the LHC. These bounds are indicated by the vertical red lines in this figure. The region ruled out by LEP is indicated by the dark grey region hatched with slanted lines. The region ruled out by the hadronic collider Tevatron is indicated by the cross-hatched region. The above mentioned red lines mark the edges of these regions giving us the pre-LHC exclusion. The region excluded by the LHC in March 2012 is indicated by light grey region marked by lines slanted in a direction opposite to the LEP exclusion region.

As one can see from this figure, before the LHC direct search constraints, the allowed mass range for the Higgs was $115 \leq M_h \lesssim 150 - 160$ and $180 \lesssim M_h \lesssim 200$ GeV. The LHC experiments ruled out existence of an SM Higgs in a major part of this range. As a result in March 2012, the mass value allowed for a SM Higgs by a combination of the EW precision measurements and 'direct' collider constraints was as indicated by the small white slit around 125 GeV. Failure to find a Higgs in this small 'allowed' mass range would then have meant the death for the SM. Indeed a new boson was found with properties very similar to a SM Higgs in precisely this mass range. This discussion should make it very clear to us that the value of the mass of the observed Higgs boson itself tested the SM at loop level to a very great accuracy.

In fact it won't be out of place to recapitulate at this point how the SM was validated and tested at various levels by discovery of new particles whose masses were predicted : either in terms of a free parameter of the model which could be determined from experiments OR 'indirectly' by comparing loop effects on physical observables with their precision measurement.

- Observation of suppression of FCNC implied that the quarks must come in isospin doublets. Thus charm was predicted since the existence of the s quark was known and top was *predicted* to be present once the b was found. Further, the very demand of cancellation of anomalies so as avoid these spoiling the renormalisability, implied existence of third generation of quarks *AND* leptons once the τ was found.
- One could get indirect information on M_c, M_t from flavour changing neutral current processes induced by loops. Agreement of this 'indirect' information with 'direct' measurements 'proved' the correctness of description of EW interactions in terms of a gauge theory.
- CP violation in meson systems could be explained in terms of the SM parameters and measured CKM mixing in quark sector *only if* three generations of quarks exist.
- M_W, M_Z was predicted in terms of $\sin\theta_W$ and direct observation of the W, Z at the predicted mass tested the particle content and tree level coupling of the matter fermions with the gauge bosons W, Z.
- Study of energy dependence of the $e^+e^- \to W^+W^-$ process gave *direct* evidence for the tree level ZWW coupling and also for the role played by this vertex in taming the bad high energy behaviour of the cross-section. So in that sense, Fig. 21 gives evidence for the gauge symmetry (ZWW coupling as indicated by symmetry) and the symmetry breaking (nonzero W mass) as well.
- Further, Teavtron found evidence for 'direct' production of the top quark at the mass M_t which was in ageeement with the value obtained 'indirectly' from precision measurement of M_W, M_Z, considering effect of radiative corrections to these masses.
- Last but not the least the existence of a minimum of $\Delta\chi^2$ at a finite nonzero mass for the SM fits

to the EW precision measurements, gave an 'indirect' proof of the existence of the Higgs. Before the 'direct' discovery of the Higgs this was also an 'indirect' probe of the couplings of the Higgs with gauge bosons and the t quarks. Further, the same fits gave an 'indirect' determination of M_h which now agrees completely with the measured mass of the observed Higgs.

Now we can turn once again to the discussion of Fig. 5. As was already indicated by the right panel of Fig. 26, the 'directly' measured value of the Higgs mass $M_h = 125.09 \pm 0.24$ GeV is right in the 'allowed' white slit and indeed confirms the SM at loop level most spectacularly. At this point, it is worth noting that if we improve upon the accuracy of measurements of M_t, M_w and M_h we can indeed hope to look for effects by loops of heavy particles which are not present in the SM but are expected to exist in various extensions of the SM, which are in turn postulated to address various shortcomings of the SM!

As already mentioned, the Higgs mass range allowed by the EW precision measurements can change when one goes away from the SM. In fact before the 'direct' discovery of the Higgs, a lot of effort had gone on, in constructing models which would allow one to avoid these constraints, should experiments reveal a Higgs boson not consistent with the bounds from the EW precision measurements. Of course, not only that many of these are not required, but some are now even ruled out, by the observation of the light state. An example of one such model is the SM with a fourth sequential generation of fermions, leptons and quarks. Since in the SM there is no guiding principle for total number of generations of fermions, except that they should be the same for quarks and leptons, this in principle is the simplest extension of the SM by addition of more matter particles to it. Observation of the low mass ~ 125 GeV scalar ruled out this extension very conclusively.

5 Observed mass of Higgs and the SM

As we saw above the EW precision measurements did put 'indirect' bounds on the Higgs mass. However, theoretically there is no information on the mass of the Higgs in the SM, as it is determined by λ an arbitrary parameter. Recall M_h and λ are related by $M_h^2 = 2\lambda v^2$. The observed mass of the Higgs determines the self coupling λ:

$$\lambda = 0.5 M_h^2/v^2 \simeq 0.13$$

This is the last free parameter of the SM that needed to be determined. Thus the only part of the scalar potential now that needs to be experimentally verified 'directly' is the triple Higgs and the quartic Higgs coupling in Eq. 56. Now that one 'knows' the value of λ one can assess the possibilities of measuring it at current and future colliders. One might ask the question whether this is the only nontrivial information about the SM that we can extract from the observed value of the mass of the Higgs. Asked differently, can one use this observed value of M_h to infer something about the SM as well as the physics beyond the SM, *viz.* the BSM. Since in these lectures we restrict ourselves to the SM, I will only talk about the possible implication of the observed Higgs mass for the SM itself.

While the SM has no 'prediction' for M_h, requirement of theoretical consistencies imply bounds on the same. These theoretical limits on the mass of the Higgs boson come from demanding good high energy behavior of scattering amplitudes in the $SU(2)_L \times U(1)_Y$ gauge theory and from the quantum corrections that the self coupling λ of Eq. 45 receives. These limits are thus essentially an artifact of the quantum field theoretical description. Let us discuss this one by one.

5.1 Unitarity bound

Recall our discussion in section 3.2.7 of the high energy behaviour of scattering amplitudes. We discussed therein the high energy behavior of the scattering amplitude $W^+W^- \to W^+W^-$. Various contributing diagrams are shown in Fig. 27. Each of these diagrams gives a contribution which grows as s^α with $\alpha = 1, 2$ where s is the centre of mass energy of the WW. This divergence appears in the

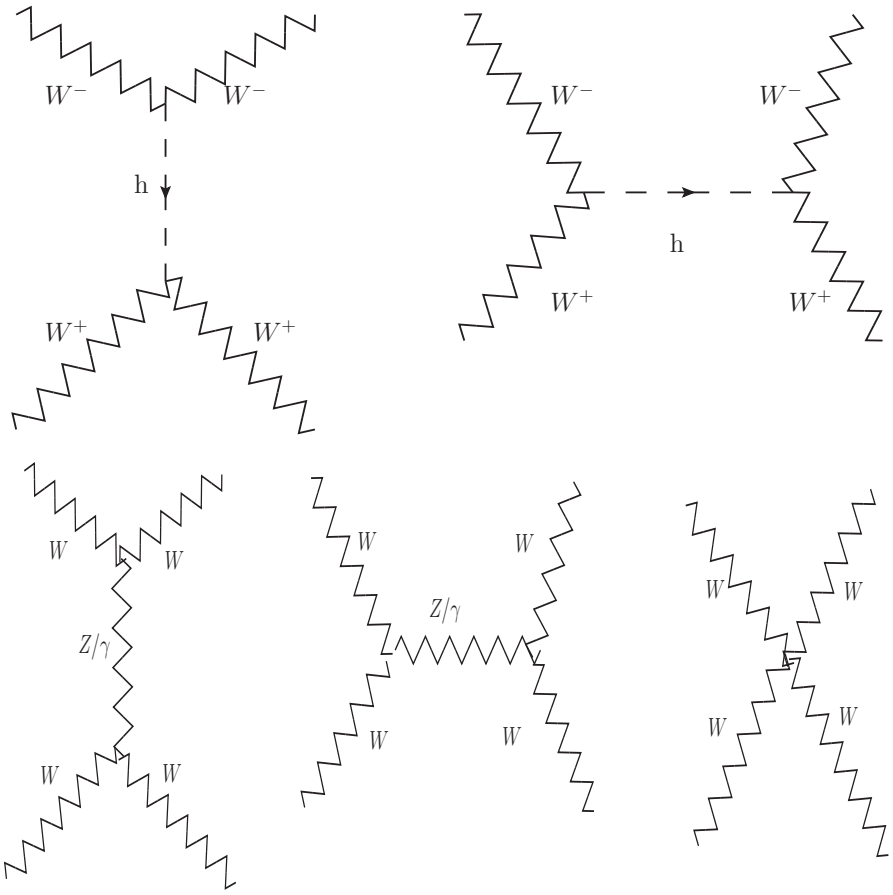


Fig. 27: The upper panel shows digrams involving h bosons contributing to $W + W \to WW$ scattering. The s-channel diagram will of course contribute only for $W^+W^- \to W^+W^-$ scattering. The lower panel shows the all the diagrams which involve exchange of the gauge bosons Z and γ as well as the one involving pure gauge vertex.

scattering of longitudinal W's. However in the SM all the divergent terms in the $WW \to WW$ amplitude cancel among each other after adding the contributions of all the diagrams shown in Fig. 27. The contribution of the h exchange diagrams as well as the that from the diagrams with pure gauge vertices play an essential role in this cancellation as mentioned before. The cancellation of the power divergences is independent of the Higgs mass and thus the requirement of non-divergent behavior does not single out any scale. Among the non divergent part of the amplitude $\mathcal{A}(WW \to WW)$, left over after all this cancellations, the contributions of the Higgs exchange diagrams shown in the top panel of Fig. 27 dominate and are dependent on the Higgs mass. These were investigated in [33] and they showed that though not divergent these can become non negligible for large values of M_h. The non-divergent part of this invariant amplitude can be written as [33]

$$\mathcal{A}(W_L^+W_L^- \to W_L^+W_L^-) = -\sqrt{2}G_\mu M_h^2 \left(\frac{s}{s-M_h^2} + \frac{t}{t-M_h^2}\right).$$

From a partial wave analysis of this amplitude one can show that this amplitude will violate tree level unitarity if

$$M_h > \left(\frac{8\pi\sqrt{2}}{3G_\mu}\right)^{1/2} \sim 1000 \text{ GeV}.$$

Thus, the theory will be strongly interacting if M_h were to exceed this value. As things stand, the observed value of M_h implies $\lambda \simeq 0.13$, far from the strongly interacting region and also safe from any unitarity violation. Thus the observed mass of the Higgs boson satisfies the unitarity bound.

5.2 Triviality and Stability bound

Effect of loop corrections to the self coupling λ in a scalar field theory, in the presence of a high scale and additional interactions of the scalar with gauge bosons and matter, was first studied decades ago [34] with an aim to examine whether one could constrain the scalar mass and other high scale masses from pure theoretical considerations. Triviality bound results from considering loop corrections to the scalar potential in Eq. 56. One demands that the quartic coupling λ in the Higgs potential from Eq. 56 reproduced below,

$$V_h = \lambda v h^3 + \lambda/4 h^4,$$

remains perturbative as well as positive at all energy scales under loop corrections. The corrections come from two sets of diagrams shown somewhat schematically in Fig. 28. The top panel shows loop

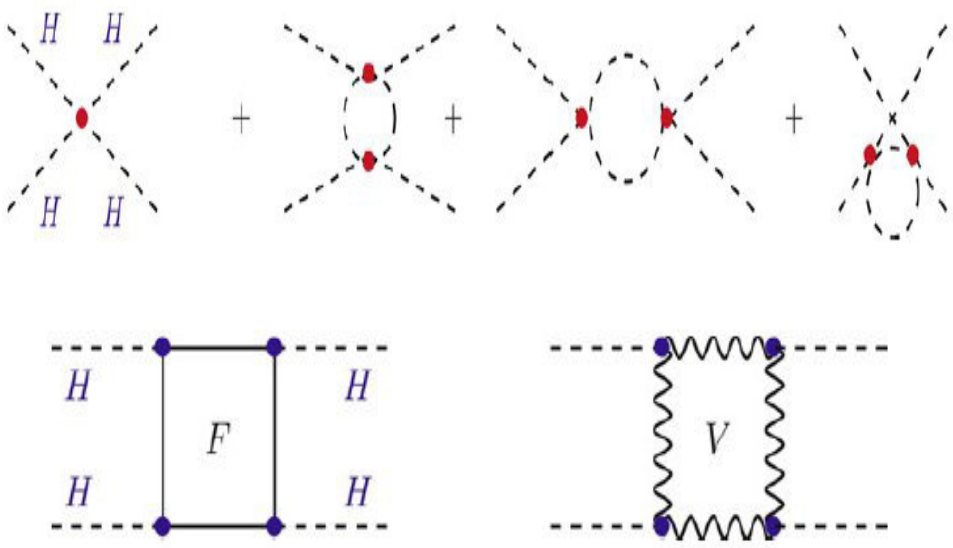


Fig. 28: The top panel shows loop corrections to the quartic coupling λ from the Higgs sector itself. The diagrams in the lower panel show contributions to the running of λ from fermion and gauge loops.

corrections to the quartic coupling λ from the Higgs sector itself whereas the diagrams in the lower panel show contributions to the running of λ from fermion and gauge loops. So the diagrams shown in the top panel are applicable to any scalar with quartic self interaction. The ones in the lower panel are specific to a gauge theory.

5.2.1 Triviality Bound

The triviality bound comes from demanding that λ should always remain perturbative. To understand the origin of this bound let us consider the case of large M_h. Since $M_h^2 = \lambda v^2$, at large m_h and hence large λ, loop corrections are dominated by the h–loops shown in the top panel of Fig. 28. A straightforward evaluation of this gives us

$$\frac{d\lambda(Q^2)}{d\log Q^2} = \frac{3}{4\pi}\lambda^2(Q^2) \tag{92}$$

Solving this, one gets

$$\lambda(Q^2) = \frac{\lambda(v^2)}{[1 - \frac{3}{4\pi^2}\lambda(v^2)\log(\frac{Q^2}{v^2})]}. \tag{93}$$

A look at Eq. 93 shows us that at large $Q^2 \gg v^2$, $\lambda(Q^2)$ can develops a pole, the so called Landau pole, at some high scale Q depending on the value λ at the EW scale v. If we demand that λ remains always in perturbative regime, then the ONLY solution would be $\lambda = 0$. This would then mean that the theory will be trivial. That of course does not make for a sensible theory. Thus the starting value of $\lambda(v)$ and hence M_h is not allowed by these considerations.

One can understand this in yet another way. If we demand that the scale at which λ blows up is above a given scale Λ, then using Eq. 93 we find that for a given value of M_h and hence $\lambda(v)$, the scale at which the Landau pole lies will be given by

$$\Lambda_C = v\ \exp\left(\frac{2\pi^2}{3\lambda}\right) = v \exp\left(\frac{4\pi^2 v^2}{3M_h^2}\right). \tag{94}$$

Thus, for example, using $\Lambda_C = \Lambda = 10^{16}$ GeV, we will find $M_h \lesssim 200$ GeV.

This bound is called the triviality bound. In simple terms it means that the value of λ at the EW scale (and hence the mass M_h) should be small enough so that $\lambda(Q^{2)}$ does not develop a pole up to a scale $Q = \Lambda_C$. Hence, if M_h were found to have a mass larger than the triviality bound, it would have meant existence of new physics below the scale Λ_C. This thus tells us that just the mass of the h can give us an indication about the scale at which SM must be complemented by additional new physics. The mass of the Higgs being only 125.09 GeV this is rather an academic discussion as this small value of the coupling λ at the EW scale, implies that the loop effects will not be driving the self coupling λ toward the Landau pole at an energy scale of interest. There are other issues that we need to address given that the observed mass is so small. But we will not discuss them here.

5.2.2 Stability bound

When M_h is small and λ is not large, the fermion/gauge boson loops are important. Even more important is that the fermions loops come with a negative sign. This means that if the fermion mass is large enough the loop corrections may drive λ negative at some scale, unless the starting value of $\lambda(v)$ is large enough. These considerations will imply a lower bound for $\lambda(v)$ and hence for M_h. This limit on M_h is called the vacuum stability bound. Now one works in the limit of small λ, opposite to the one used when considering the triviality bound. Hence the contribution of the h-loops shown in the upper panel of Fig. 28 can be neglected. Hence the equation for energy dependence of λ now can be written as:

$$\frac{d\lambda(Q^2)}{d\log(Q^2)} \simeq \frac{1}{16\pi^2}[12\lambda^2 + 6\lambda f_t^2 - \mathbf{3f_t^4} - \frac{\mathbf{3}}{\mathbf{2}}\lambda(\mathbf{3g_2^2} + \mathbf{g_1^2}) + \frac{3}{16}(2g_2^4 + (g_2^2 + g_1^2)^2)] \tag{95}$$

$\mathbf{f_t} = \frac{\sqrt{\mathbf{2}}\mathbf{M_t}}{\mathbf{v}}$ is the Yukawa coupling for the top. Since $M_t \sim 173$ GeV and $v \simeq 246$ GeV, one can see that the Yukawa coupling is $\simeq 1$. Thus it will dominate the scale dependence of λ. At small M_h and hence small $\lambda(v)$, λ can turn negative at some value of Q. Recall the Higgs potential. A negative value of λ will mean an unbounded potential and clearly the vacuum will be unstable. The condition for non negativity of λ and hence vacuum stability, is

$$M_h^2 > \frac{v^2}{8\pi^2}\log(Q^2/v^2)\left[12m_t^2/v^4 - \frac{3}{16}(2g_2^4 + (g_2^2 + g_1^2)^2)\right]. \tag{96}$$

Again, depending upon the scale up to which we demand the potential to be positive definite, we find that the starting value $\lambda(v)$ (and hence M_h) has to be above a critical value dependent on the scale. If

we demand that the $\lambda(Q)$ is positive up to Λ_C we then get a lower bound on M_h. For example choosing, $\Lambda_C = 10^3 GeV$ we get $M_h \gtrsim 70$ GeV. This bound is called the stability bound.

In the above analysis we have demanded that $\lambda(\Lambda)$ does not become negative so that the potential is stable. This is the condition for absolute stability of vacuum. However, Planck scale dynamics might stabilise the vacuum for $|\Phi| >> v$ and we might be living in a metastable vacuum which has a life time bigger than that of the Universe. The cartoon shown in Fig. 29 indicates such a situation. One can then obtain lower bounds on M_h demanding that vacuum is metastable with a life time bigger than the life time of the Universe. Clearly evaluation of these bounds can not be presented in the simplistic analysis that we have given here.

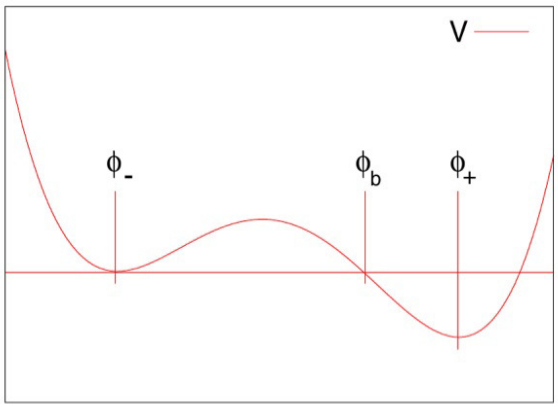


Fig. 29: Cartoon of a field configuration that would give rise to metastable vacuum.

A complete and sophisticated analysis of Ref. [35] in fact gives the vacuum stability bounds on the Higgs mass taking into account the effect of renormalisation group evolution(RGE) as well as that of metastability of the vacuum. Fig. 30 taken from Ref. [35] shows the stability bounds, indicated by the pale yellow green area, as a function of scale at which the instability sets in. The spread is due to the theoretical uncertainties, major ones being the top mass uncertainty and the missing higher order contributions to the equations. RGE takes into account not just the one loop corrections shown in Fig. 28 but also includes the resummation of leading logarithmic corrections. As one can can see even from the simple minded analysis presented here, the bound depends critically on the value of f_t and hence on M_t. If one overlays the bounds on the Higgs mass of Fig. 26 obtained 'indirectly' from the EW precision analysis as well as the LEP/Teavtron/LHC searches then we realise that the thin white silver which was still allowed by March 2013 corresponds to the boundary of the pale yellow-green region indicating the stability bound. Due to the finite width of these bands caused by various uncertainties mentioned above, the observed mass of the Higgs M_h may or may not be consistent with the hypothesis that the SM remains consistent all the way to Planck scale. Given that everything depends logarithmically on different scales and with the high accuracy of the experimental measurement of M_h, the need to do the evolution of λ taking into account higher order effects is thus clear.

In fact the need for more accurate calculation was already apparent, even before the Higgs discovery, with the rather low values of M_h indicated by the 'indirect' limits. To appreciate this, look at Fig. 26 again disregarding the vertical red lines corresponding to the LHC 95% bound, which delineate the pale grey region hatched with inclined lines. The 3σ region around the minimum of $\Delta\chi^2$ and hence preferred by the EW precision data, allowed by Tevatron data, $115 \leq M_h \leq 150$ GeV, covers the range of masses where the stability bound is operative and the upper limits on the possible scale of new physics indicated by the vacuum (in)stability interesting. The need for accuracy in the theoretical prediction of stability

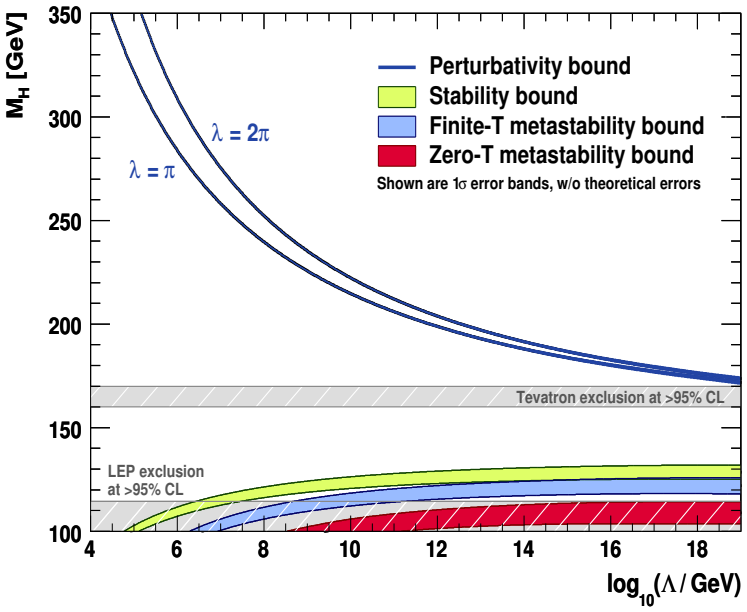


Fig. 30: The vacuum stability bound on M_h as a function of the scale. Bounds are shown for absolute stability as well as metastability. Taken from [35].

bound is thus very apparent. In May 2012, with the discovery of the Higgs imminent, an NNLO analysis of the problem became available [36], which reduced the theoretical error on the bounds coming from the unknown higher order corrections to ~ 1 GeV.

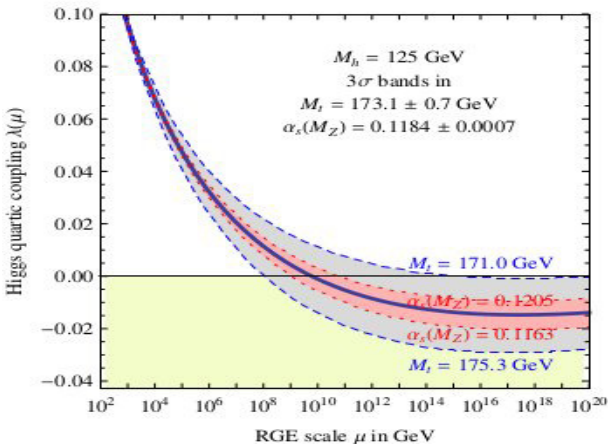


Fig. 31: $\lambda(\mu)$ as a function of scale for different values of α_s, M_t varied within the experimental errors. The plot is taken from [36].

However, there still remains a sizable error due to the errors in experimentally determined parameters M_t, α_s. Fig. 31, taken from [36], shows behavior of $\lambda(\mu)$ as a function of the energy scale μ. One now sees clearly that the scale at which λ becomes zero and hence the vacuum unstable, depends critically on M_t and the strong coupling α_s. For example, for the central value of M_t used, μ value at which λ becomes zero changes by at least an order of magnitude as α_s is varied within errors. The dependence on M_t is even stronger. We will comment later on the range of M_t used in this analysis. According to

this analysis the absolute stability of the vacuum up to Planck scale M_{pl} is guaranteed for,

$$M_h\,[\,\text{GeV}] > 129.4 + 1.4\left(\frac{M_t\,[\text{GeV}] - 173.1}{0.7}\right) - 0.5\left(\frac{\alpha_s(M_Z) - 0.1184}{0.0007}\right) \pm 1.0\,_{\text{th}}. \qquad (97)$$

In this analysis the error on pole mass of the top was taken to be $\Delta m_t = \pm 0.7$ GeV. Taking into account the errors, Eq. 97 then means that for $m_h < 126$ GeV, vacuum stability of the SM all the way to Planck Scale is excluded at 98% c.l. Clearly, this value is far too close to the observed value of 125.09 ± 0.24 GeV to require careful considerations of various issues before we draw conclusions about the validity of the SM at high scale. For the measured value of the Higgs mass, the exact scale where λ crosses zero, though not M_{pl} seems close to it and depends entirely on the exact value of M_t and M_h. Indeed these considerations may be relevant for consideration of BSM or models of inflation etc.

The same can be seen clearly from Fig. 32 taken from [36]. This shows the results of this NNLO analysis of the region in M_h–M_t plane from the vacuum stability considerations. The left panel shows

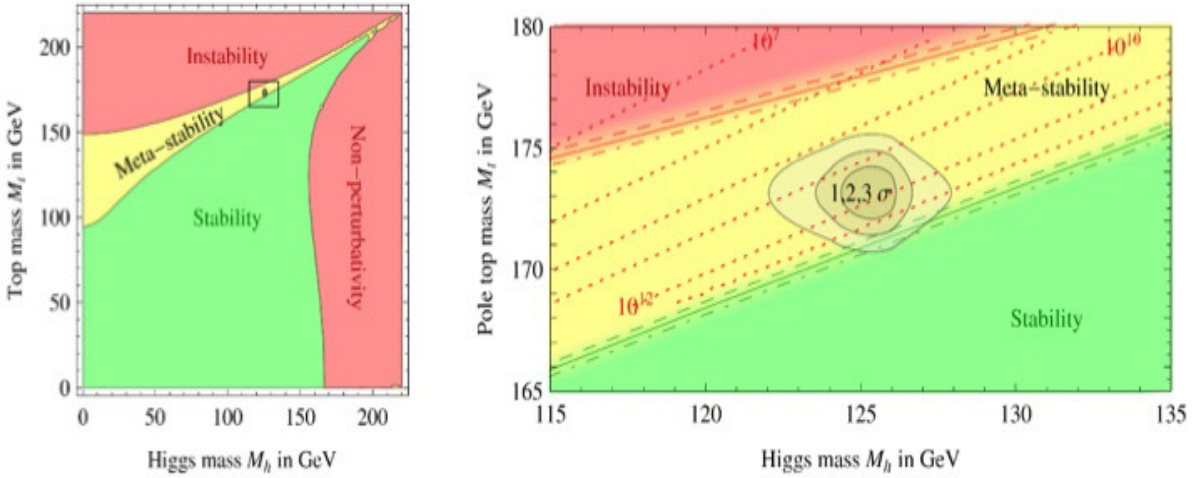


Fig. 32: The left panel shows the regions in the M_t–M_h plane where the vacuum is absolutely stable, metastable and unstable. Right panel shows the zoom-in the region of values preferred experimentally. The grey areas show allowed regions at 1,2 and 3 σ. The three curves on the boundary of two regions correspond to three values of α_s. Superimposed on it are the contours of constant value of the high scale where the instability occurs. The plot is taken from [36]

the regions in the M_t–M_h plane where the vacuum is absolutely stable, metastable and unstable. To understand the role and size of various 'experimental' uncertainties the right panel shows a zoom in of the region around the experimentally determined M_h–M_t values. The grey areas show allowed regions at 1,2 and 3 σ. The three curves on the boundary of two regions correspond to three values of α_s. Superimposed on it are the contours of constant value of the high scale where the instability occurs. We see that the experimentally determined values lie right on the boundary of the stable/metastable region. The answer to the question as to whether or not, the experimentally determined value of M_h (known now to a high accuracy $M_h = 125.09 \pm 0.24$ GeV) is consistent with SM vacuum being (meta)stable all the way to Planck scale, very much depends on M_t values.

Let us discuss this issue in a little more detail. The stability bounds given in [36] used errors on m_t as measured at the hadronic colliders the Tevatron and the LHC. This is the so called Monte Carlo or kinematic mass, which is a parameter in the Monte Carlos used while analysing the data and studying

the top quark production at the colliders. Conversion of this parameter into the pole mass, which is the parameter required in these theoretical considerations and for the RGE, has uncertainties coming from hadronisation and fragmentation models, underlying event etc. These are typically non perturbative in character. Another way to extract the pole mass in a well defined manner is to extract $M_t^{\overline{MS}}$, the mass of the top quark in the $\overline{MS}$ scheme from the measurement of the top quark cross-sections at the Tevatron and the NNLO calculation of the same. The procedure to convert this mass to the pole mass $M_t(M_t)$, leads to uncertainties in M_t larger than the 0.7 GeV taken in Eq. 97. This exercise, using the available information in 2012 led to an estimate of the pole mass for the top [37]:

$$M_t^{pole} = 173.3 \pm 2.8 \text{ GeV}.$$

Compare this with the error of 0.7 GeV that was used in the estimate obtained in [36]. The vacuum stability constraint now becomes $M_h > 129.4 \pm 5.6$ GeV instead of the one in Eq. 97. This observation then can weaken the conclusion about the high scale upto which the SM remains valid without getting into conflict with stability. The future International Linear Collider(ILC) can measure the top mass M_t to a high accuracy of 100 MeV. What is more important is the fact that the determination of the t mass

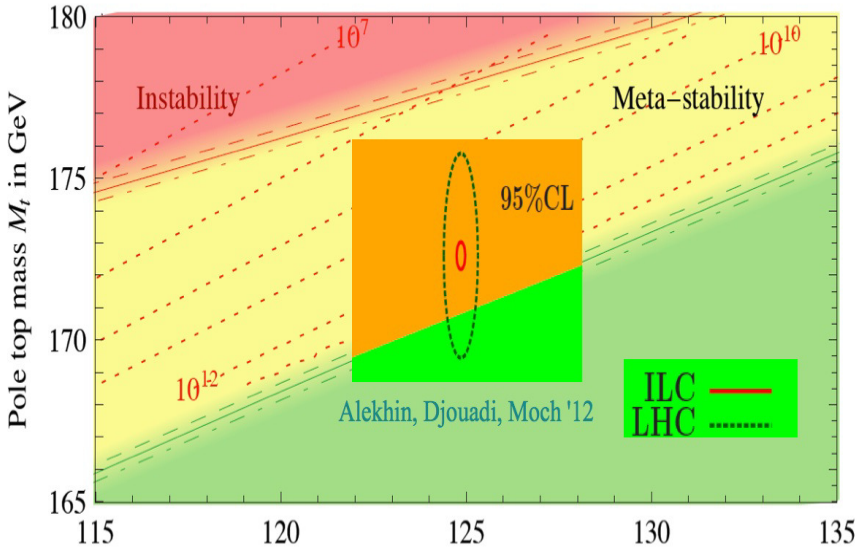


Fig. 33: This is the same figure as in the right panel of Fig. 32, where the zoomed region around experimentally determined values from [36] has ben overlaid with the uncertainties of M_t determination as extracted in [37]. This was done by G. Isidori in his talk at SUSY 2014.

at the ILC comes directly from measurement of the $t\bar{t}$ production cross-section in e^+e^- collisions, near the $t\bar{t}$ threshold. This can be measured very accurately and has been computed theoretically to a high precision as well. This measurement can be converted into the pole mass in an unambiguous way. Fig. 33 shows how such a precision measurement of the mass at the ILC can really shed light on whether the currently measured higgs mass points to the NEED of BSM physics at any **particular high scale**. In the above figure, the bigger blue circle has been drawn assuming an LHC accuracy of t mass measurement of 1 GeV. However, a reduction of this error to about 500 MeV looks possible and is an active area of research. These kind of investigations are just the next logical step in our efforts to test the SM through a combination of the 'direct' and 'indirect' observations.

6 Concluding remarks

In any case the days of Standard Model are coming to an end in some sense! Hopefully it will be the case of 'The King is Dead' and 'Long live the King'! We have, however, not much idea what particular BSM option, if any, would be the new king. As we have discussed above, already the mass of the observed state can be used to answer the question about the scale upto which the SM is valid. In fact, this has been one of the most impressive facts about the SM. It has held the ability to ask and answer questions about its own consistency within its structure. Just like the **gauge principle** and the **unitarity** were the guiding principle so far, now the **small** mass of the discovered Higgs ($\sim \mathcal{O}$ weak scale) might be the guiding principle for future theoretical developments! This will be discussed in other lectures at the school. We should get a peek at the BSM land through the 'window' of measurement of the properties of the Higgs and the top quark! Exciting days are ahead for sure! If 14 TeV LHC should also fail to find

Fig. 34: The Higgs and Top portal for BSM physics.

'direct' evidence for the BSM physics we would really have to understand what is so special about the Standard Model. Precision measurements of the observed Higgs mass and Higgs couplings will be then our window to this world of physics beyond the SM.

7 Acknowledgements

I would like to acknowledge an ongoing collaboration with Sunil Mukhi to write a book on the Standard Model which is in preparation. I will also like to express my appreciation of the patience shown by Martijn Mulders in waiting for my lecture notes. I acknowledge hospitality of the Abdus Salam International Centre for Theoretical Physics where a substantial part of these notes was completed. Financial support from the Department of Science and Technology, India under Grant No. SR/S2/JCB-64/2007 under the J.C. Bose Fellowship scheme is gratefully acknowledged. Last but not the least, I would like to acknowledge the AEPSHEP school organisers for asking me to give these lectures and the students for listening with patience! I would also like to acknowledge help from Ms. Anuja Thakar for providing some of the artwork and Dr. Gaurav Mendiratta for providing the drawing for Fig. 9.

8 About References

In the bibliography I have listed the original theory papers and the early experimental papers which are referred to in the text from time to time. After that I list a large number of very good text books where one can find detailed discussions of many of the issues involved.

References

[1] G. Aad *et al.* [ATLAS Collaboration], Phys. Lett. B **716** (2012) 1 [arXiv:1207.7214 [hep-ex]]; S. Chatrchyan *et al.* [CMS Collaboration], Phys. Lett. B **716** (2012) 30 [arXiv:1207.7235 [hep-ex]].

[2] G. Aad *et al.* [ATLAS and CMS Collaborations], Phys. Rev. Lett. **114** (2015) 191803 [arXiv:1503.07589 [hep-ex]].

[3] K. A. Olive *et al.* [Particle Data Group Collaboration], Chin. Phys. C **38** (2014) 090001.

[4] P. W. Higgs, Phys. Lett. **12** (1964) 132; P. W. Higgs, Phys. Rev. Lett. **13** (1964) 508. F. Englert and R. Brout, Phys. Rev. Lett. **13** (1964) 321.

[5] E. Fermi, Z. Phys. **88** (1934) 161.

[6] E. C. G. Sudarshan and R. e. Marshak, Phys. Rev. **109** (1958) 1860.

[7] R. P. Feynman and M. Gell-Mann, Phys. Rev. **109** (1958) 193.

[8] N. Cabibbo, Phys. Rev. Lett. **10** (1963) 531. doi:10.1103/PhysRevLett.10.531

[9] M. Kobayashi and T. Maskawa, Prog. Theor. Phys. **49** (1973) 652.

[10] H. Albrecht *et al.* [ARGUS Collaboration], Phys. Lett. B **192** (1987) 245.

[11] S. L. Glashow, Nucl. Phys. **22** (1961) 579, S. Weinberg, Phys. Rev. Lett. **19** (1967) 1264. A. Salam, "Weak and Electromagnetic Interactions," Conf. Proc. C **680519** (1968) 367.

[12] F. Abe *et al.* [CDF Collaboration], Phys. Rev. D **50** (1994) 2966. F. Abe *et al.* [CDF Collaboration], Phys. Rev. Lett. **73** (1994) 225, [hep-ex/9405005], S. Abachi *et al.* [D0 Collaboration], Phys. Rev. Lett. **74** (1995) 2422 [hep-ex/9411001], F. Abe *et al.* [CDF Collaboration], Phys. Rev. Lett. **74** (1995) 2626 [hep-ex/9503002], S. Abachi *et al.* [D0 Collaboration], Phys. Rev. Lett. **74** (1995) 2632 [hep-ex/9503003].

[13] C. Quigg, "Top-ology," Phys. Today **50N5** (1997) 20, hep-ph/9704332.

[14] The url for the Gfitter group giving results for the global electroweak standard model fit is http://project-gfitter.web.cern.ch/project-gfitter/$\mathrm{Standard\_Model}$/

[15] "Horizons in Physics", Editor A.W. Joshi, John Wiley Publication (1989).

[16] Y. Nambu and G. Jona-Lasinio, Phys. Rev. **122** (1961) 345; Y. Nambu and G. Jona-Lasinio, Phys. Rev. **124** (1961) 246.

[17] S. L. Glashow, J. Iliopoulos and L. Maiani, Phys. Rev. D **2** (1970) 1285.

[18] M. K. Gaillard and B. W. Lee, Phys. Rev. D **10** (1974) 897, M. K. Gaillard, B. W. Lee and J. L. Rosner, Rev. Mod. Phys. **47** (1975) 277, M. K. Gaillard, B. W. Lee and R. E. Shrock, Phys. Rev. D **13** (1976) 2674.

[19] J. S. Bell, Nucl. Phys. B **60** (1973) 427.

[20] C. H. Llewellyn Smith, Phys. Lett. **46B** (1973) 233, S. D. Joglekar, Annals Phys. **83** (1974) 427, J. M. Cornwall, D. N. Levin and G. Tiktopoulos, Phys. Rev. D **10** (1974) 1145 Erratum: [Phys. Rev. D **11** (1975) 972].

[21] G. 't Hooft, Nucl. Phys. B **35** (1971) 167; G. 't Hooft and M. J. G. Veltman, Nucl. Phys. B **44** (1972) 189.

[22] F. J. Hasert *et al.* [Gargamelle Neutrino Collaboration], Phys. Lett. B **46** (1973) 138.

[23] F. J. Hasert *et al.*, Phys. Lett. B **46** (1973) 121.

[24] D. P. Barber *et al.* [MARK-J Collaboration], Phys. Rev. Lett. **46** (1981) 1663.

[25] J. E. Kim, P. Langacker, M. Levine and H. H. Williams, Rev. Mod. Phys. **53** (1981) 211.

[26] G. Arnison *et al.* [UA1 Collaboration], Phys. Lett. B **122**, 103 (1983); G. Arnison *et al.* [UA1 Collaboration], Phys. Lett. B **126** (1983) 398.

[27] M. Banner *et al.* [UA2 Collaboration], Phys. Lett. B **122** (1983) 476; P. Bagnaia *et al.* [UA2 Collaboration], Phys. Lett. B **129** (1983) 130.

[28] S. Schael *et al.* [ALEPH and DELPHI and L3 and OPAL and LEP Electroweak Collaborations], Phys. Rept. **532** (2013) 119, [arXiv:1302.3415 [hep-ex]].

[29] A. Sirlin, Phys. Rev. D **22** (1980) 971; A. Sirlin and W. J. Marciano, Nucl. Phys. B **189** (1981) 442; C. H. Llewellyn Smith and J. F. Wheater, Phys. Lett. B **105** (1981) 486.

[30] M. J. G. Veltman, Acta Phys. Polon. B **8** (1977) 475; Acta Phys. Polon. B **25** (1994) 1627.

[31] S. Schael *et al.* [ALEPH and DELPHI and L3 and OPAL and SLD and LEP Electroweak Working Group and SLD Electroweak Group and SLD Heavy Flavour Group Collaborations], Phys. Rept. **427** (2006) 257

[32] The url of the LEP and Tevatron Electroweak Working Group is http://lepewwg.web.cern.ch/LEPEWWG/

[33] B. W. Lee, C. Quigg and H. B. Thacker, Phys. Rev. Lett. **38** (1977) 883; B. W. Lee, C. Quigg and H. B. Thacker, Phys. Rev. D **16** (1977) 1519.

[34] N. Cabibbo, L. Maiani, G. Parisi and R. Petronzio, Nucl. Phys. B **158** (1979) 295.

[35] J. Ellis, J. R. Espinosa, G. F. Giudice, A. Hoecker and A. Riotto, Phys. Lett. B **679** (2009) 369 [arXiv:0906.0954 [hep-ph]].

[36] G. Degrassi, S. Di Vita, J. Elias-Miro, J. R. Espinosa, G. F. Giudice, G. Isidori and A. Strumia, JHEP **1208** (2012) 098 [arXiv:1205.6497 [hep-ph]].

[37] S. Alekhin, A. Djouadi and S. Moch, Phys. Lett. B **716** (2012) 214, [arXiv:1207.0980 [hep-ph]].

[38] D.H. Perkins, Introduction to High Energy Physics, Cambridge University Press, 4th Edition, 2000.

[39] T.P. Cheng and Ling-Fong Li, 'Gauge Theory of Elementary Particle Physics', Clarendon Press - Oxford, 1982.

[40] F. Halzen and A. Martin, 'Quarks and Leptons : An Introductory Course in Modern Particle Physics', John Wiley and Sons, 1984.

[41] M. Peskin and S. Schröder, 'An Introduction to Quantum Field Theory', Perseus Books Publishing, 1995.

[42] W. Greiner and B. Mueller, 'Gauge Theories of Weak Interactions', Third Edition, Springer Verlag, 2000.

[43] C. Burgess and G. Moore, 'The Standard Model : a primer', Cambridge University Press, 2006.

[44] C. Quigg, 'Gauge theories of the strong, weak and electromagnetic interactions, Princeton University Press, Second Edition, 2013.

[45] R.M. Godbole and S. Mukhi, 'Standard Model of Particle Physics' (To be published by Cambridge University Press).

Proceedings of the 2014 Asia–Europe–Pacific School of High-Energy Physics, Puri, India, 4–17 November 2014, edited by
M. Mulders and R. Godbole, CERN Yellow Reports: School Proceedings, Vol. 2/2017, CERN-2017-005-SP (CERN, Geneva, 2017)

QCD*

P. Z. Skands
Theoretical Physics, CERN, Geneva, Switzerland
and
School of Physics & Astronomy, Monash University, Clayton, Australia

Abstract
These lecture notes are directed at a level suitable for graduate students in High Energy Physics. They are intended to give an introduction to the theory and phenomenology of quantum chromodynamics (QCD), focusing on collider-physics applications. The aim is to bring the reader to a level where informed decisions can be made concerning different approaches and their uncertainties. The material is divided into five main areas: (1) fundamentals, (2) fixed-order perturbative QCD, (3) Monte Carlo event generators and parton showers, (4) Matching at Leading and Next-to-Leading Order, and (5) Soft QCD physics.

Keywords
Lectures; quantum chromodynamics; gauge theory; jet: fragmentation; Monte Carlo; PYTHIA.

Useful complementary references

List of additional general study/reference material:

- basic quantum field theory: Ref. [1],
- textbooks on QCD: Refs. [2, 3],
- jets and jet algorithms: Ref. [4],
- general-purpose event generators: Ref. [5],
- the string model: Ref. [6],
- step-by-step PYTHIA tutorial: see 'worksheet' available on the PYTHIA homepage,
- Monte Carlo methods and random numbers: Refs. [7, 8].

1 Introduction

When probed at very short wavelengths, quantum chromodynamics (QCD) is essentially a theory of free 'partons'—quarks and gluons—which only scatter off one another through relatively small quantum corrections that can be systematically calculated. At longer wavelengths, of the order of the size of the proton $\sim 1\,\mathrm{fm} = 10^{-15}\,\mathrm{m}$. However, we see strongly bound towers of hadron resonances emerge, with string-like potentials building up if we try to separate their partonic constituents. Due to our inability to perform analytic calculations in strongly coupled field theories, QCD is, therefore, still only partially solved. Nonetheless, all its features, across all distance scales, are believed to be encoded in a single one-line formula of alluring simplicity: the Lagrangian[1] of QCD.

*Originally based on lectures given at ESHEP 2010 (Raseborg, Finland) and subsequently updated for TASI 2012 (Boulder, Colorado) and AEPSHEP 2014 (Puri, India).

[1]Throughout these notes we let it be implicit that 'Lagrangian' really refers to Lagrangian density, $\mathcal{L}$, the four-dimensional space–time integral of which is the action.

https://doi.org/10.23730/CYRSP-2017-002.63

The consequence for collider physics is that some parts of QCD can be calculated in terms of the fundamental parameters of the Lagrangian, whereas others must be expressed through models or functions whose effective parameters are not a priori calculable, but which can be constrained by fits to data. However, even in the absence of a perturbative expansion, there are still several strong theorems which hold, and which can be used to give relations between seemingly different processes. (This is, e.g., the reason it makes sense to constrain parton-distribution functions (PDFs) in ep collisions and then re-use the same ones for pp collisions.) Thus, in the sections dealing with phenomenological models, we shall emphasize that the loss of a factorized perturbative expansion is not equivalent to a total loss of predictivity.

An alternative approach would be to give up on calculating QCD and use leptons instead. Formally, this amounts to summing inclusively over strong-interaction phenomena, when such are present. While such a strategy might succeed in replacing what we do know about QCD by 'unity', even the most adamant chromophobe must acknowledge a few basic facts of collider physics for the next decade(s). (1) At the Large Hadron Collider (LHC), the initial states are hadrons, and hence, at the very least, well-understood and precise PDFs will be required. (2) High precision will mandate calculations to higher orders in perturbation theory, which, in turn, will involve more QCD. (3) The requirement of lepton *isolation* makes the very definition of a lepton depend implicitly on QCD. (4) The rate of jets that are misreconstructed as leptons in the experiment depends explicitly on QCD. (5) Finally, although many new-physics signals *do* give observable signals in the lepton sector, this is far from guaranteed, nor is it exclusive when it occurs. It would, therefore, be unwise not to attempt to solve QCD to the best of our ability, the better to prepare ourselves for both the largest possible discovery reach and the highest attainable subsequent precision.

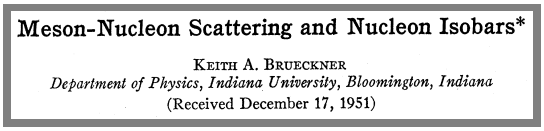
Moreover, QCD is the richest gauge theory we have, so far, encountered. Its emergent phenomena, unitarity properties, colour structure, non-perturbative dynamics, quantum versus classical limits, interplay between scale-invariant and scale-dependent properties and its wide range of phenomenological applications are still very much topics of active investigation, about which we continue to learn.

In addition, or perhaps as a consequence, the field of QCD is currently experiencing something of a revolution. On the perturbative side, new methods to compute scattering amplitudes with very high particle multiplicities are being developed, together with advanced techniques for combining such amplitudes with all-orders resummation frameworks. On the non-perturbative side, the wealth of data on soft-physics processes from the LHC is forcing us to reconsider the reliability of the standard fragmentation models, and heavy-ion collisions are providing new insights into the collective behaviour of hadronic matter. The study of cosmic rays impinging on the Earth's atmosphere challenges our ability to extrapolate fragmentation models from collider energy scales to the region of ultra-high-energy cosmic rays. And finally, dark-matter annihilation processes in space may produce hadrons, whose spectra are sensitive to the modelling of fragmentation.

In the following, we shall focus on QCD for mainstream collider physics. This includes the basics of the gauge group SU(3), colour factors, the running of $\alpha_{\rm s}$, factorization, hard processes, IR safety, parton showers and matching, event generators, hadronization and the so-called underlying event. While not covering everything, hopefully these topics can also serve at least as stepping stones to more specialized issues that have been left out, such as twistor-inspired techniques, heavy flavours, polarization or forward physics, or to topics more tangential to other fields, such as axions, lattice QCD or heavy-ion physics.

1.1 A first hint of colour

Looking for new physics, as we do now at the LHC, it is instructive to consider the story of the discovery of colour. The first hint was arguably the Δ^{++} baryon, discovered in 1951 [9]. The title and part of the abstract from this historical paper are reproduced in Fig. 1. In the context of the quark model—which first had to be developed, successively joining together the notions of spin, isospin, strangeness, and the



Meson-Nucleon Scattering and Nucleon Isobars*

KEITH A. BRUECKNER
Department of Physics, Indiana University, Bloomington, Indiana
(Received December 17, 1951)

"[...] It is concluded that the apparently anomalous features of the scattering can be interpreted to be an indication of a resonant meson-nucleon interaction corresponding to a nucleon isobar with spin $\frac{3}{2}$, isotopic spin $\frac{3}{2}$, and with an excitation energy of 277 MeV."

Fig. 1: The title and part of the abstract of the 1951 paper [9] (published in 1952) in which the existence of the Δ^{++} baryon was deduced, based on data from Sachs and Steinberger at Columbia [10] and from Anderson *et al.* at Chicago [11]. Further studies at Chicago were quickly performed in Ref. [12, 13]. See also the memoir by Nagle [14].

eightfold way[2]—the flavour and spin content of the Δ^{++} baryon is

$$\left|\Delta^{++}\right\rangle = |\, u_\uparrow\, u_\uparrow\, u_\uparrow\rangle\,, \tag{1}$$

clearly a highly symmetric configuration. However, since the Δ^{++} is a fermion, it must have an overall antisymmetric wave function. In 1965, 14 years after its discovery, this was finally understood by the introduction of colour as a new quantum number associated with the group SU(3) [15, 16]. The Δ^{++} wave function can now be made antisymmetric by arranging its three quarks antisymmetrically in this new degree of freedom,

$$\left|\Delta^{++}\right\rangle = \epsilon^{ijk}\,|\, u_{i\uparrow}\, u_{j\uparrow}\, u_{k\uparrow}\rangle\,, \tag{2}$$

and hence solving the mystery.

More direct experimental tests of the number of colours were provided first by measurements of the decay width of $\pi^0 \to \gamma\gamma$ decays, which is proportional to $N_{\rm C}^2$, and later by the famous 'R' ratio in $\mathrm{e^+e^-}$ collisions ($R = \sigma(\mathrm{e^+e^-} \to \mathrm{q\bar{q}})/\sigma(\mathrm{e^+e^-} \to \mu^+\mu^-)$), which is proportional to $N_{\rm C}$, see, e.g., Ref. [3]. Below, in Section 1.2 we shall see how to calculate such colour factors.

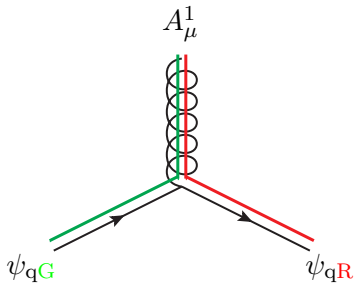
1.2 The Lagrangian of QCD

QCD is based on the gauge group $\mathrm{SU}(3)$, the Special Unitary group in three (complex) dimensions, whose elements are the set of unitary 3×3 matrices with determinant one. Since there are nine linearly independent unitary complex matrices[3], one of which has determinant -1, there are a total of eight independent directions in this matrix space, corresponding to eight different generators as compared with the single one of quantum electrodynamics (QED). In the context of QCD, we normally represent this group using the so-called *fundamental*, or *defining*, representation, in which the generators of $\mathrm{SU}(3)$ appear as a set of eight traceless and hermitean matrices, to which we return below. We shall refer to indices enumerating the rows and columns of these matrices (from 1 to 3) as *fundamental* indices, and we use the letters i, j, k, ..., to denote them. We refer to indices enumerating the generators (from 1 to 8), as *adjoint* indices[4], and we use the first letters of the alphabet (a, b, c, ...) to denote them. These matrices can operate both on each other (representing combinations of successive gauge transformations) and on a set of 3-vectors, the latter of which represent quarks in colour space; the quarks are *triplets* under $\mathrm{SU}(3)$. The matrices can be thought of as representing gluons in colour space (or, more precisely, the gauge

[2] In physics, the 'eightfold way' refers to the classification of the lowest-lying pseudoscalar mesons and spin-$1/2$ baryons within octets in SU(3)-flavour space (u, d, s). The Δ^{++} is part of a spin-$3/2$ baryon decuplet, a 'tenfold way' in this terminology.

[3] A complex $N \times N$ matrix has $2N^2$ degrees of freedom, on which unitarity provides N^2 constraints.

[4] The dimension of the *adjoint*, or *vector*, representation is equal to the number of generators, $N^2 - 1 = 8$ for $\mathrm{SU}(3)$, while the dimension of the fundamental representation is the degree of the group, $N = 3$ for $\mathrm{SU}(3)$.



$$\propto \; -\tfrac{i}{2} g_{\mathrm{s}} \quad \bar{\psi}_{\mathrm{qR}} \quad \lambda^1 \quad \psi_{\mathrm{qG}}$$

$$= \; -\tfrac{i}{2} g_{\mathrm{s}} \quad \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Fig. 2: Illustration of a qqg vertex in QCD, before summing/averaging over colours: a gluon in a state represented by λ^1 interacts with quarks in the states ψ_{qR} and ψ_{qG}.

transformations carried out by gluons), and hence there are eight different gluons; the gluons are *octets* under $\mathrm{SU}(3)$.

The Lagrangian density of QCD is

$$\mathcal{L} = \bar{\psi}^i_{\mathrm{q}} (i\gamma^\mu)(D_\mu)_{ij} \psi^j_{\mathrm{q}} - m_{\mathrm{q}} \bar{\psi}^i_{\mathrm{q}} \psi_{\mathrm{q}i} - \frac{1}{4} F^a_{\mu\nu} F^{a\mu\nu}, \tag{3}$$

where ψ^i_{q} denotes a quark field with (fundamental) colour index i, $\psi_{\mathrm{q}} = (\psi_{\mathrm{qR}}, \psi_{\mathrm{qG}}, \psi_{\mathrm{qB}})^{\mathrm{T}}$, γ^μ is a Dirac matrix that expresses the vector nature of the strong interaction, with μ being a Lorentz vector index, m_{q} allows for the possibility of non-zero quark masses (induced by the standard Higgs mechanism or similar), $F^a_{\mu\nu}$ is the gluon field strength tensor for a gluon with (adjoint) colour index a (i.e., $a \in [1, \ldots, 8]$) and D_μ is the covariant derivative in QCD,

$$(D_\mu)_{ij} = \delta_{ij} \partial_\mu - i g_{\mathrm{s}} t^a_{ij} A^a_\mu \,, \tag{4}$$

with g_{s} the strong coupling (related to α_{s} by $g^2_{\mathrm{s}} = 4\pi\alpha_{\mathrm{s}}$; we return to the strong coupling in more detail below) A^a_μ the gluon field with colour index a, and t^a_{ij} proportional to the hermitean and traceless Gell–Mann matrices of $\mathrm{SU}(3)$,

$$\lambda^1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; \lambda^2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; \lambda^3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; \lambda^4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

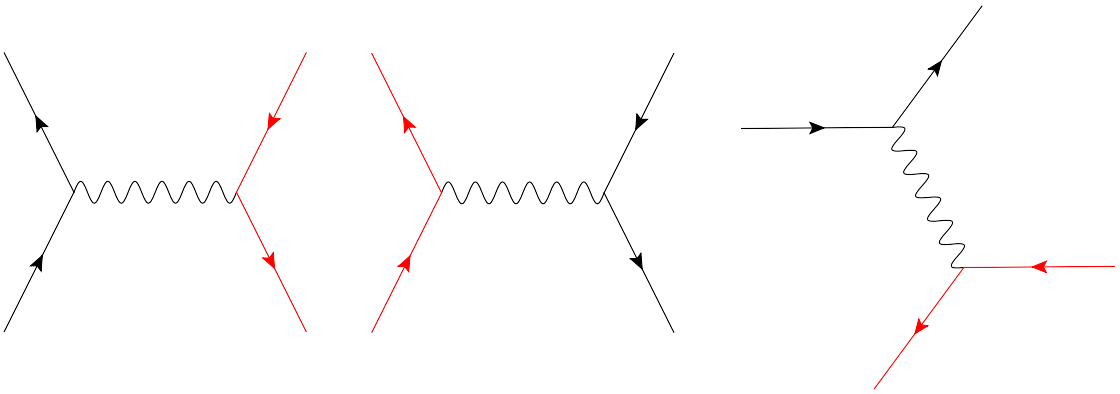
$$\lambda^5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \; \lambda^6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \; \lambda^7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \; \lambda^8 = \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & \frac{-2}{\sqrt{3}} \end{pmatrix}. \tag{5}$$

These generators are just the $\mathrm{SU}(3)$ analogues of the Pauli matrices in $\mathrm{SU}(2)$. By convention, the constant of proportionality is normally taken to be

$$t^a_{ij} = \frac{1}{2} \lambda^a_{ij}. \tag{6}$$

This choice, in turn, determines the normalization of the coupling g_{s}, via Eq. (4), and fixes the values of the $\mathrm{SU}(3)$ Casimirs and structure constants, to which we return below.

An example of the colour flow for a quark–gluon interaction in colour space is given in Fig. 2. Normally, of course, we sum over all the colour indices, so this example merely gives a pictorial representation of what one particular (non-zero) term in the colour sum looks like.



Hadronic Z decay	Drell–Yan	DIS
$\mathrm{e}^-\mathrm{e}^+ \to \gamma^*/Z^0 \to \mathrm{q}\bar{\mathrm{q}}$	$\mathrm{q}\bar{\mathrm{q}} \to \gamma^*/Z^0 \to \ell^+\ell^-$	$\ell\bar{\mathrm{q}} \overset{\gamma^*/Z^*}{\to} \ell\bar{\mathrm{q}}$
$\propto N_\mathrm{C}$	$\propto 1/N_\mathrm{C}$	$\propto 1$

Fig. 3: Illustration of the three crossings of the interaction of a lepton current (black) with a quark current (red) via an intermediate photon or Z boson, with corresponding colour factors.

1.3 Colour factors

Typically, we do not measure colour in the final state—instead we average over all possible incoming colours and sum over all possible outgoing ones, wherefore QCD scattering amplitudes (squared), in practice, always contain sums over quark fields contracted with Gell–Mann matrices. These contractions in turn produce traces which yield the *colour factors* that are associated with each QCD process, and which basically count the number of 'paths through colour space' that the process at hand can take[5].

A very simple example of a colour factor is given by the decay process $\mathrm{Z} \to \mathrm{q}\bar{\mathrm{q}}$. This vertex contains a simple δ_{ij} in colour space; the outgoing quark and antiquark must have identical (anti-)colours. Squaring the corresponding ME and summing over final-state colours yields a colour factor of

$$\mathrm{e}^+\mathrm{e}^- \to \mathrm{Z} \to \mathrm{q}\bar{\mathrm{q}} \quad : \quad \sum_{\text{colours}} |M|^2 \propto \delta_{ij}\delta^*_{ji} = \mathrm{Tr}\{\delta\} = N_\mathrm{C} = 3, \tag{7}$$

since i and j are quark (i.e., three-dimensional fundamental) indices.

A next-to-simplest example is given by $\mathrm{q}\bar{\mathrm{q}} \to \gamma^*/\mathrm{Z} \to \ell^+\ell^-$ (usually referred to as the Drell–Yan process [17]), which is just a crossing of the previous one. By crossing symmetry, the squared ME, including the colour factor, is exactly the same as before, but since the quarks are, here, incoming, we must *average* rather than sum over their colours, leading to

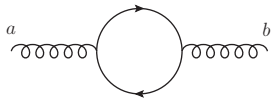
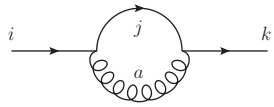
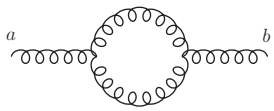
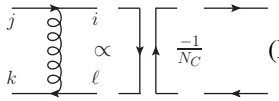
$$\mathrm{q}\bar{\mathrm{q}} \to \mathrm{Z} \to \mathrm{e}^+\mathrm{e}^- \quad : \quad \frac{1}{9}\sum_{\text{colours}} |M|^2 \propto \frac{1}{9}\delta_{ij}\delta^*_{ji} = \frac{1}{9}\mathrm{Tr}\{\delta\} = \frac{1}{3}, \tag{8}$$

where the colour factor now expresses a *suppression* which can be interpreted as due to the fact that only quarks of matching colours are able to collide and produce a Z boson. The chance that a quark and an antiquark picked at random from the colliding hadrons have matching colours is $1/N_\mathrm{C}$.

Similarly, $\ell\mathrm{q} \to \ell\mathrm{q}$ via t-channel photon exchange (usually called deep inelastic scattering—DIS—with 'deep' referring to a large virtuality of the exchanged photon), constitutes yet another crossing of the same basic process, see Fig. 3. The colour factor in this case comes out as unity.

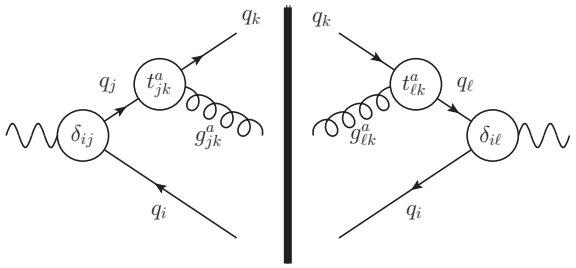
[5]The convention choice represented by Eq. (6) introduces a 'spurious' factor of two for each power of the coupling α_s. Although one could, in principle, absorb that factor into a redefinition of the coupling, effectively redefining the normalization of 'unit colour charge', the standard definition of α_s is now so entrenched that alternative choices would be counter-productive, at least in the context of a pedagogical review.

Table 1: Trace relations for t matrices (convention-independent). More relations can be found in Ref. [2, Section 1.2] and in Ref. [1, Appendix A.3].

Trace relation	Indices	Occurs in diagram squared
$\mathrm{Tr}\{t^a t^b\} = T_R\,\delta^{ab}$	$a, b \in [1, \ldots, 8]$	
$\sum_a t^a_{ij} t^a_{jk} = C_F\,\delta_{ik}$	$a \in [1, \ldots, 8]$ $i, j, k \in [1, \ldots, 3]$	
$\sum_{c,d} f^{acd} f^{bcd} = C_A\,\delta^{ab}$	$a, b, c, d \in [1, \ldots, 8]$	
$t^a_{ij} t^a_{k\ell} = T_R \left(\delta_{jk}\delta_{i\ell} - \frac{1}{N_C}\delta_{ij}\delta_{k\ell}\right)$	$i, j, k, \ell \in [1, \ldots, 3]$	 (Fierz)

To illustrate what happens when we insert (and sum over) quark–gluon vertices, such as the one depicted in Fig. 2, we take the process $Z \to 3$ jets. The colour factor for this process can be computed as follows, with the accompanying illustration showing a corresponding diagram (squared) with explicit colour-space indices on each vertex:

$$
\begin{aligned}
&\mathrm{Z} \to \mathrm{q g \bar{q}} \;: \\
&\sum_{\text{colours}} |M|^2 \;\propto\; \delta_{ij} t^a_{jk}\,(t^a_{\ell k}\delta^*_{i\ell})^* \\
&\qquad\qquad = \; \mathrm{Tr}\{t^a t^a\} \\
&\qquad\qquad = \; \frac{1}{2}\mathrm{Tr}\{\delta\} = 4,
\end{aligned}
\qquad (9)
$$



where the last $\mathrm{Tr}\{\delta\} = 8$, since the trace runs over the eight-dimensional adjoint indices.

The tedious task of taking traces over t matrices can be greatly alleviated by use of the relations given in Table 1. In the standard normalization convention for the $\mathrm{SU}(3)$ generators, Eq. (6), the Casimirs of $\mathrm{SU}(3)$ appearing in Table 1 are[6]

$$
T_R = \frac{1}{2} \qquad C_F = \frac{4}{3} \qquad C_A = N_\mathrm{C} = 3. \qquad (10)
$$

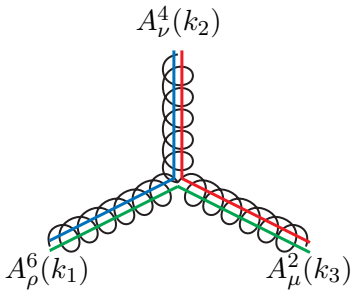
In addition, the gluon self-coupling on the third line in Table 1 involves factors of f^{abc}. These are called the *structure constants* of QCD and they enter via the non-Abelian term in the gluon-field-strength tensor appearing in Eq. (3),

$$
F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g_\mathrm{s} f^{abc} A^b_\mu A^c_\nu. \qquad (11)
$$

The structure constants of $\mathrm{SU}(3)$ are listed in the box below. They define the *adjoint*, or *vector*, representation of $\mathrm{SU}(3)$ and are related to the fundamental-representation generators via the commutator relations

$$
t^a t^b - t^b t^a = [t^a, t^b] = \mathrm{i} f^{abc} t_c, \qquad (12)
$$

[6] See, e.g., Ref. [1, Appendix A.3] for how to obtain the Casimirs in other normalization conventions.



$$\propto \; -g_\mathrm{s}\, f^{246} \quad \begin{array}{l}[(k_3-k_2)^\rho g^{\mu\nu} \\ +(k_2-k_1)^\mu g^{\nu\rho} \\ +(k_1-k_3)^\nu g^{\rho\mu}]\end{array}$$

Fig. 4: Illustration of a ggg vertex in QCD, before summing/averaging over colours: interaction between gluons in the states λ^2, λ^4, and λ^6 is represented by the structure constant f^{246}.

or equivalently,

$$\mathrm{i}f^{abc} \;=\; 2\mathrm{Tr}\{t^c[t^a,t^b]\}. \tag{13}$$

Thus, it is a matter of choice whether one prefers to express colour space on a basis of fundamental-representation t matrices, or via the structure constants f, and one can go back and forth between the two.

Structure constants of SU(3)

$$f_{123} = 1 \tag{14}$$

$$f_{147} = f_{246} = f_{257} = f_{345} = \frac{1}{2} \tag{15}$$

$$f_{156} = f_{367} = -\frac{1}{2} \tag{16}$$

$$f_{458} = f_{678} = \frac{\sqrt{3}}{2} \tag{17}$$

Antisymmetric in all indices

All other $f_{abc} = 0$

Expanding the $F_{\mu\nu}F^{\mu\nu}$ term of the Lagrangian using Eq. (11), we see that there is a 3-gluon and a 4-gluon vertex that involve f^{abc}, the latter of which has two powers of f and two powers of the coupling.

Finally, the last line of Table 1 is not really a trace relation but instead a useful so-called Fierz transformation, which expresses products of t matrices in terms of Kronecker δ functions. It is often used, for instance, in shower Monte Carlo applications, to assist in mapping between colour flows in $N_\mathrm{C} = 3$, in which cross sections and splitting probabilities are calculated, and those in $N_\mathrm{C} \to \infty$ ('leading colour') are used to represent colour flow in the Monte Carlo (MC) 'event record'.

A gluon self-interaction vertex is illustrated in Fig. 4, to be compared with the quark–gluon interaction in Fig. 2. We remind the reader that gauge boson self-interactions are a hallmark of non-Abelian theories and that their presence leads to some of the main differences between QED and QCD. One should also keep in mind that the colour factor for the vertex in Fig. 4, C_A, is roughly twice as large as that for a quark, C_F.

1.4 The strong coupling

To first approximation, QCD is *scale invariant*. That is, if one 'zooms in' on a QCD jet, one will find a repeated self-similar pattern of jets within jets within jets, reminiscent of fractals. In the context of QCD, this property was originally called light-cone scaling, or Bjørken scaling. This type of scaling

is closely related to the class of angle-preserving symmetries, called *conformal* symmetries. In physics today, the terms 'conformal' and 'scale invariant' are used interchangeably[7]. Conformal invariance is a mathematical property of several QCD-'like' theories which are now being studied (such as $N = 4$ supersymmetric relatives of QCD). It is also related to the physics of so-called 'unparticles', although that is a relation that goes beyond the scope of these lectures.

Regardless of the labelling, if the strong coupling did not run (we shall return to the running of the coupling below), Bjørken scaling would be absolutely true. QCD would be a theory with a fixed coupling, the same at all scales. This simplified picture already captures some of the most important properties of QCD, as we shall discuss presently.

In the limit of exact Bjørken scaling—QCD at fixed coupling—properties of high-energy interactions are determined only by *dimensionless* kinematic quantities, such as scattering angles (pseudorapidities) and ratios of energy scales[8]. For applications of QCD to high-energy collider physics, an important consequence of Bjørken scaling is, thus, that the rate of bremsstrahlung jets, with a given transverse momentum, scales in direct proportion to the hardness of the fundamental partonic-scattering process they are produced in association with. This agrees well with our intuition about accelerated charges; the harder you 'kick' them, the harder the radiation they produce.

For instance, in the limit of exact scaling, a measurement of the rate of 10 GeV jets produced in association with an ordinary Z boson could be used as a direct prediction of the rate of 100 GeV jets that would be produced in association with a 900 GeV Z′ boson, and so forth. Our intuition about how many bremsstrahlung jets a given type of process is likely to have should, therefore, be governed, first and foremost, by the *ratios* of scales that appear in that particular process, as has been highlighted in a number of studies focusing on the mass and $p_\perp$ scales appearing, for example, in beyond-the-standard-model (BSM) physics processes [18–21]. Bjørken scaling is also fundamental to the understanding of jet substructure in QCD, see, for example, Refs. [22, 23].

On top of the underlying scaling behaviour, the running coupling will introduce a dependence on the absolute scale, implying more radiation at low scales than at high ones. The running is logarithmic with energy and is governed by the so-called *beta function*,

$$Q^2 \frac{\partial \alpha_\mathrm{s}}{\partial Q^2} = \frac{\partial \alpha_\mathrm{s}}{\partial \ln Q^2} = \beta(\alpha_\mathrm{s}), \tag{18}$$

where the function driving the energy dependence, the beta function, is defined as

$$\beta(\alpha_\mathrm{s}) = -\alpha_\mathrm{s}^2(b_0 + b_1\alpha_\mathrm{s} + b_2\alpha_\mathrm{s}^2 + \cdots), \tag{19}$$

with LO (one-loop) and NLO (two-loop) coefficients

$$b_0 = \frac{11C_A - 4T_R n_\mathrm{f}}{12\pi}, \tag{20}$$

$$b_1 = \frac{17C_A^2 - 10T_R C_A n_\mathrm{f} - 6T_R C_F n_\mathrm{f}}{24\pi^2} = \frac{153 - 19n_\mathrm{f}}{24\pi^2}. \tag{21}$$

In the b_0 coefficient, the first term is due to gluon loops while the second is due to quark loops. Similarly, the first term of the b_1 coefficient arises from double gluon loops, while the second and third represent mixed quark–gluon loops. At higher loop orders, the b_i coefficients depend explicitly on the renormalization scheme that is used. A brief discussion can be found in the Particle Data Group (PDG) review on

[7] Strictly speaking, conformal symmetry is more restrictive than just scale invariance, but examples of scale-invariant field theories that are not conformal are rare.

[8] Originally, the observed approximate agreement with this was used as a powerful argument for pointlike substructure in hadrons; since measurements at different energies are sensitive to different resolution scales, independence of the absolute energy scale is indicative of the absence of other fundamental scales in the problem and hence of pointlike constituents.

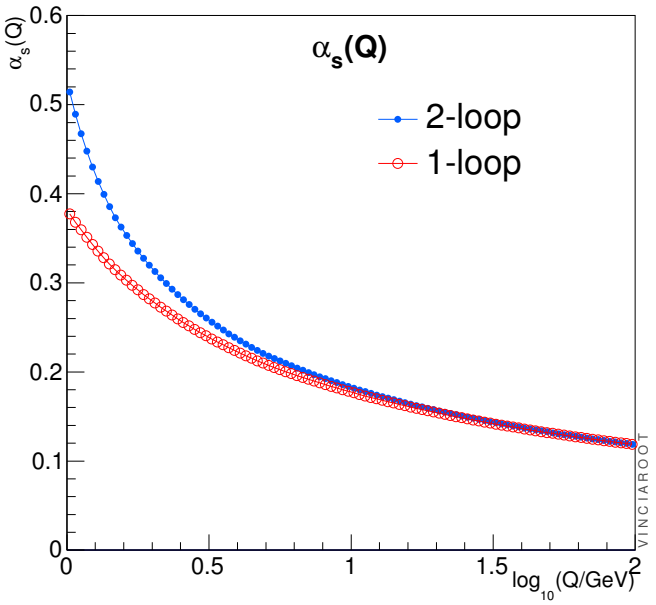


Fig. 5: Illustration of the running of α_s at one- (open circles) and two-loop order (filled circles), starting from the same value of $\alpha_s(M_Z) = 0.12$.

QCD [24], with more elaborate ones contained in Refs. [2, 3]. Note that, if there are additional coloured particles beyond the standard-model ones, loops involving those particles enter at energy scales above the masses of the new particles, thus modifying the running of the coupling at high scales. This is discussed, for example, for supersymmetric models in Ref. [25].

Numerically, the value of the strong coupling is usually specified by giving its value at the specific reference scale $Q^2 = M_Z^2$, from which we can obtain its value at any other scale by solving Eq. (18),

$$\alpha_s(Q^2) = \alpha_s(M_Z^2)\frac{1}{1 + b_0\alpha_s(M_Z^2)\ln\frac{Q^2}{M_Z^2} + \mathcal{O}(\alpha_s^2)}, \tag{22}$$

with relations including the $\mathcal{O}(\alpha_s^2)$ terms available, for example, in Ref. [2]. Relations between scales not involving M_Z^2 can obviously be obtained by just replacing M_Z^2 by some other scale Q'^2 everywhere in Eq. (22). A comparison of running at one- and two-loop order, in both cases starting from $\alpha_s(M_Z) = 0.12$, is given in Fig. 5. As is evident from the figure, the two-loop running is somewhat faster than the two-loop.

As an application, let us prove that the logarithmic running of the coupling implies that an intrinsically multi-scale problem can be converted to a single-scale one, up to corrections suppressed by two powers of α_s, by taking the geometric mean of the scales involved. This follows from expanding an arbitrary product of individual α_s factors around an arbitrary scale μ, using Eq. (22),

$$\begin{aligned}\alpha_s(\mu_1)\alpha_s(\mu_2)\cdots\alpha_s(\mu_n) &= \prod_{i=1}^{n}\alpha_s(\mu)\left(1 + b_0\,\alpha_s\ln\left(\frac{\mu^2}{\mu_i^2}\right) + \mathcal{O}(\alpha_s^2)\right)\\ &= \alpha_s^n(\mu)\left(1 + b_0\,\alpha_s\ln\left(\frac{\mu^{2n}}{\mu_1^2\mu_2^2\cdots\mu_n^2}\right) + \mathcal{O}(\alpha_s^2)\right),\end{aligned} \tag{23}$$

whereby the specific single-scale choice $\mu^n = \mu_1\mu_2\cdots\mu_n$ (the geometric mean) can be seen to push the difference between the two sides of the equation one order higher than would be the case for any other

combination of scales[9].

The appearance of the number of flavours, $n_{\rm f}$, in b_0 implies that the slope of the running depends on the number of contributing flavours. Since full QCD is best approximated by $n_{\rm f} = 3$ below the charm threshold, by $n_{\rm f} = 4$ and 5 from there to the b and t thresholds, respectively, and then by $n_{\rm f} = 6$ at scales higher than $m_{\rm t}$, it is, therefore, important to be aware that the running changes slope across quark-flavour thresholds. Likewise, it would change across the threshold for any coloured new-physics particles that might exist, with a magnitude depending on the particles' colour and spin quantum numbers.

The negative overall sign of Eq. (19), combined with the fact that $b_0 > 0$ (for $n_{\rm f} \leq 16$), leads to the famous result[10] that the QCD coupling effectively *decreases* with energy, called asymptotic freedom, for the discovery of which the Nobel prize in physics was awarded to D. Gross, H. Politzer and F. Wilczek in 2004. An extract of the prize announcement runs as follows.

> *What this year's Laureates discovered was something that, at first sight, seemed completely contradictory. The interpretation of their mathematical result was that the closer the quarks are to each other, the* weaker *is the 'colour charge'. When the quarks are really close to each other, the force is so weak that they behave almost as free particles*[a]*. This phenomenon is called 'asymptotic freedom'. The converse is true when the quarks move apart: the force becomes stronger when the distance increases*[b]*.*
>
> [a] More correctly, it is the coupling rather than the force which becomes weak as the distance decreases. The $1/r^2$ Coulomb singularity of the force is only dampened, not removed, by the diminishing coupling.
>
> [b] More correctly, it is the potential which grows, linearly, while the force becomes constant.

Among the consequences of asymptotic freedom is that perturbation theory becomes better behaved at higher absolute energies, due to the effectively decreasing coupling. Perturbative calculations for our 900 GeV Z′ boson from before should, therefore, be slightly faster converging than equivalent calculations for the 90 GeV one. Furthermore, since the running of $\alpha_{\rm s}$ explicitly breaks Bjørken scaling, we also expect to see small changes in jet shapes and in jet-production ratios as we vary the energy. For instance, since high-$p_\perp$ jets start out with a smaller effective coupling, their intrinsic shape (irrespective of boost effects) is somewhat narrower than for low-$p_\perp$ jets, an issue which can be important for jet calibration. Our current understanding of the running of the QCD coupling is summarized by the plot in Fig. 6, taken from a recent comprehensive review by S. Bethke [26, 27].

As a final remark on asymptotic freedom, note that the decreasing value of the strong coupling with energy must eventually cause it to become comparable to the electromagnetic and weak ones, at some energy scale. Beyond that point, which may lie at energies of order 10^{15}–10^{17} GeV (although it may be lower if as-yet-undiscovered particles generate large corrections to the running), we do not know what the further evolution of the combined theory will actually look like, or whether it will continue to exhibit asymptotic freedom.

Now consider what happens when we run the coupling in the other direction, towards smaller energies. Taken at face value, the numerical value of the coupling diverges rapidly at scales below 1 GeV, as illustrated by the curves disappearing off the left-hand edge of the plot in Fig. 6. To make this divergence explicit, one can rewrite Eq. (22) in the following form,

$$\alpha_{\rm s}(Q^2) = \frac{1}{b_0 \ln \frac{Q^2}{\Lambda^2}}, \tag{24}$$

where

$$\Lambda \sim 200\,\text{MeV} \tag{25}$$

[9] In a fixed-order calculation, the individual scales μ_i, would correspond, for example, to the n hardest scales appearing in an IR-safe sequential-clustering algorithm applied to the given momentum configuration.

[10] Perhaps the highest pinnacle of fame for Eq. (19) was reached when the sign of it featured in an episode of the TV series 'Big Bang Theory'.

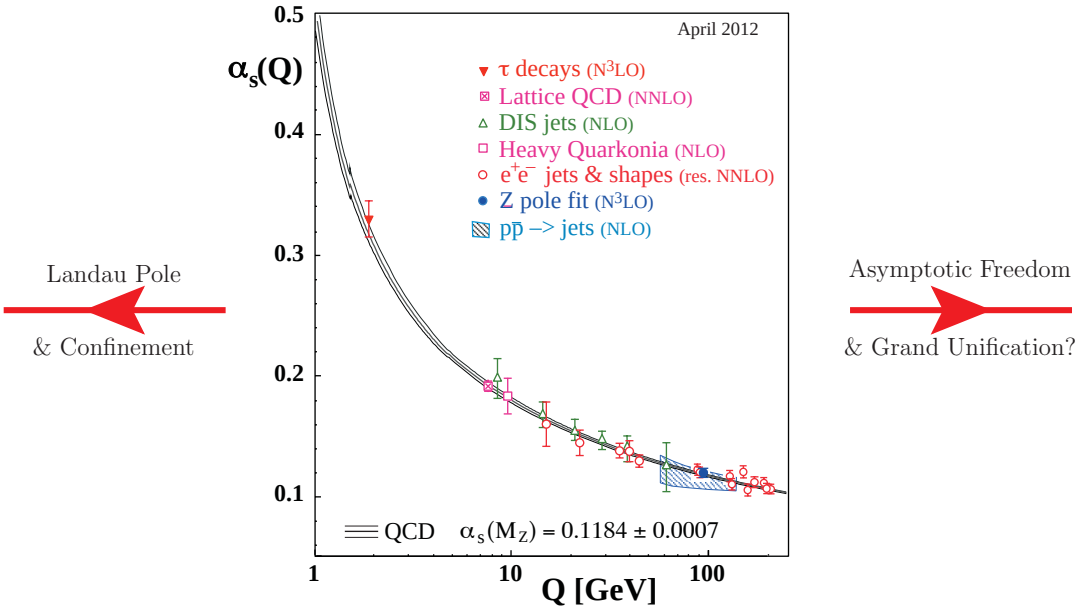


Fig. 6: Illustration of the running of α_s in a theoretical calculation (band) and in physical processes at different characteristic scales, from Refs. [24, 26]. The little kinks at $Q = m_c$ and $Q = m_b$ are caused by discontinuities in the running across the flavour thresholds.

specifies the energy scale at which the perturbative coupling would nominally become infinite, called the Landau pole. (Note, however, that this only parameterizes the purely *perturbative* result, which is not reliable at strong coupling, so Eq. (24) should not be taken to imply that the physical behaviour of full QCD should exhibit a divergence for $Q \to \Lambda$.)

Finally, one should be aware that there is a multitude of different ways of defining both Λ and $\alpha_s(M_Z)$. At the very least, the numerical value one obtains depends both on the renormalization scheme used (with the dimensional-regularization-based 'modified minimal subtraction' scheme, $\overline{\text{MS}}$, being the most common one) and on the perturbative order of the calculations used to extract them. As a rule of thumb, fits to experimental data typically yield smaller values for $\alpha_s(M_Z)$ the higher the order of the calculation used to extract it (see, e.g., Refs. [24, 26–28]), with $\alpha_s(M_Z)|_{\text{LO}} \gtrsim \alpha_s(M_Z)|_{\text{NLO}} \gtrsim \alpha_s(M_Z)|_{\text{NNLO}}$. Further, since the number of flavours changes the slope of the running, the location of the Landau pole for fixed $\alpha_s(M_Z)$ depends explicitly on the number of flavours used in the running. Thus, each value of n_f is associated with its own value of Λ, with the following matching relations across thresholds guaranteeing continuity of the coupling at one loop,

$$n_f = 5 \leftrightarrow 6 \quad : \quad \Lambda_6 = \Lambda_5 \left(\frac{\Lambda_5}{m_t}\right)^{\frac{2}{21}} \qquad \Lambda_5 = \Lambda_6 \left(\frac{m_t}{\Lambda_6}\right)^{\frac{2}{23}} , \tag{26}$$

$$n_f = 4 \leftrightarrow 5 \quad : \quad \Lambda_5 = \Lambda_4 \left(\frac{\Lambda_4}{m_b}\right)^{\frac{2}{23}} \qquad \Lambda_4 = \Lambda_5 \left(\frac{m_b}{\Lambda_5}\right)^{\frac{2}{25}} , \tag{27}$$

$$n_f = 3 \leftrightarrow 4 \quad : \quad \Lambda_4 = \Lambda_3 \left(\frac{\Lambda_3}{m_c}\right)^{\frac{2}{25}} \qquad \Lambda_3 = \Lambda_4 \left(\frac{m_c}{\Lambda_4}\right)^{\frac{2}{27}} . \tag{28}$$

It is sometimes stated that QCD only has a single free parameter, the strong coupling. However, even in the perturbative region, the beta function depends explicitly on the number of quark flavours, as we have seen, and thereby also on the quark masses. Furthermore, in the non-perturbative region around or below Λ_{QCD}, the value of the perturbative coupling, as obtained, for example, from Eq. (24),

gives little or no insight into the behaviour of the full theory. Instead, universal functions (such as parton densities, form factors, fragmentation functions, etc), effective theories (such as the operator product expansion, chiral perturbation theory, or heavy quark effective theory), or phenomenological models (such as Regge theory or the string and cluster hadronization models) must be used, which in turn depend on additional non-perturbative parameters whose relation to, for example, $\alpha_{\mathrm{s}}(M_{\mathrm{Z}})$, is not a priori known.

For some of these questions, such as hadron masses, lattice QCD can furnish important additional insight, but for multi-scale and/or time-evolution problems, the applicability of lattice methods is still severely restricted; the lattice formulation of QCD requires a Wick rotation to Euclidean space. The time-coordinate can then be treated on an equal footing with the other dimensions, but intrinsically Minkowskian problems, such as the time evolution of a system, are inaccessible. The limited size of current lattices also severely constrain the scale hierarchies that it is possible to 'fit' between the lattice spacing and the lattice size.

1.5 Colour states

A final example of the application of the underlying $\mathrm{SU}(3)$ group theory to QCD is given by considering which colour states we can obtain by combinations of quarks and gluons. The simplest example of this is the combination of a quark and antiquark. We can form a total of nine different colour–anticolour combinations, which fall into two irreducible representations of $\mathrm{SU}(3)$:

$$3 \otimes \overline{3} = 8 \oplus 1. \tag{29}$$

The singlet corresponds to the symmetric wave function $\frac{1}{\sqrt{3}}\left(\left|R\bar{R}\right\rangle + \left|G\bar{G}\right\rangle + \left|B\bar{B}\right\rangle\right)$, which is invariant under $\mathrm{SU}(3)$ transformations (the definition of a singlet). The other eight linearly independent combinations (which can be represented by one for each Gell–Mann matrix, with the singlet corresponding to the identity matrix) transform into each other under $\mathrm{SU}(3)$. Thus, although we sometimes talk about colour-singlet states as being made up, for example, of 'red–antired', that is not quite precise language. The actual state $\left|R\bar{R}\right\rangle$ is *not* a pure colour singlet. Although it does have a non-zero *projection* onto the singlet wave function above, it also has non-zero projections onto the two members of the octet that correspond to the diagonal Gell–Mann matrices. Intuitively, one can also easily realize this by noting that an $\mathrm{SU}(3)$ rotation of $\left|R\bar{R}\right\rangle$ would in general turn it into a different state, say $\left|B\bar{B}\right\rangle$, whereas a true colour singlet would be invariant. Finally, we can also realize from Eq. (29) that a random (colour-uncorrelated) quark–antiquark pair has a $1/N^2 = 1/9$ chance to be in an overall colour-singlet state; otherwise it is in an octet.

Similarly, there are also nine possible quark–quark (or antiquark–antiquark) combinations, six of which are symmetric under interchange of the two quarks and three of which are antisymmetric:

$$6 \;=\; \begin{pmatrix} |RR\rangle \\ |GG\rangle \\ |BB\rangle \\ \frac{1}{\sqrt{2}}\left(|RG\rangle + |GR\rangle\right) \\ \frac{1}{\sqrt{2}}\left(|GB\rangle + |BG\rangle\right) \\ \frac{1}{\sqrt{2}}\left(|BR\rangle + |RB\rangle\right) \end{pmatrix}, \qquad \bar{3} = \begin{pmatrix} \frac{1}{\sqrt{2}}\left(|RG\rangle - |GR\rangle\right) \\ \frac{1}{\sqrt{2}}\left(|GB\rangle - |BG\rangle\right) \\ \frac{1}{\sqrt{2}}\left(|BR\rangle - |RB\rangle\right) \end{pmatrix}. \tag{30}$$

The members of the sextet transform into (linear combinations of) each other under $\mathrm{SU}(3)$ transformations, and similarly for the members of the antitriplet, and hence neither of these can be reduced further. The breakdown into irreducible $\mathrm{SU}(3)$ multiplets is, therefore,

$$3 \otimes 3 = 6 \oplus \overline{3}. \tag{31}$$

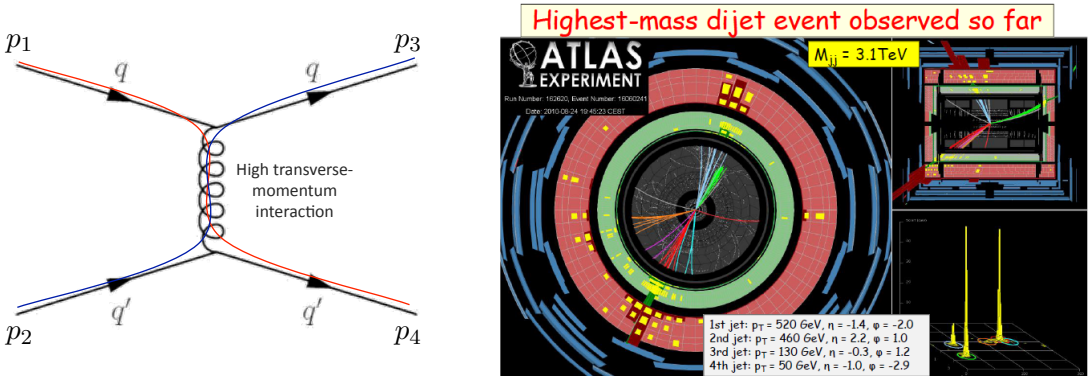


Fig. 7: *Left:* Rutherford scattering of quarks in QCD, exemplifying the type of process that dominates the short-distance interaction cross section at hadron colliders. *Right:* an example of what such a reaction looks like in a detector, in this case the ATLAS experiment.

Thus, an uncorrelated pair of quarks has a $1/3$ chance to add to an overall anti-triplet state (corresponding to coherent superpositions like '$\mathrm{red} + \mathrm{green} = \mathrm{antiblue}$'[11]); otherwise it is in an overall sextet state.

Note that the emphasis on the quark–(anti)quark pair being *uncorrelated* is important; production processes that correlate the produced partons, like $\mathrm{Z} \to \mathrm{q\bar{q}}$ or $\mathrm{g} \to \mathrm{q\bar{q}}$, will project out specific components (here the singlet and octet, respectively). Note also that, if the quark and (anti)quark are on opposite sides of the universe (i.e., living in two different hadrons), the QCD *dynamics* will not care what overall colour state they are in, so for the formation of multi-partonic states in QCD, obviously the spatial part of the wave functions (causality at the very least) will also play a role. Here, we are considering *only* the colour part of the wave functions. Some additional examples are

$$8 \otimes 8 = 27 \oplus 10 \oplus \overline{10} \oplus 8 \oplus 8 \oplus 1, \tag{32}$$

$$3 \otimes 8 = 15 \oplus 6 \oplus 3, \tag{33}$$

$$3 \otimes 6 = 10 \oplus 8, \tag{34}$$

$$3 \otimes 3 \otimes 3 = (6 \oplus \overline{3}) \otimes 3 = 10 \oplus 8 \oplus 8 \oplus 1. \tag{35}$$

Physically, the '27' in the first line corresponds to a completely incoherent addition of the colour charges of two gluons; the decuplets are slightly more coherent (with a lower total colour charge), the octets yet more, and the singlet corresponds to the combination of two gluons that have precisely equal and opposite colour charges, so that their total charge is zero. Further extensions and generalizations of these combination rules can be obtained, for example, using the method of the Young tableaux [29, 30].

2 Hard processes

Our main tool for solving QCD at high energy scales, $Q \gg \Lambda_{\mathrm{QCD}}$, is perturbative quantum field theory, the starting point for which is matrix elements (MEs) which can be calculated systematically at fixed orders in the strong coupling α_{s}. At least at the lowest order (LO), the procedure is standard textbook material [1] and it has also by now been highly automated, by the advent of tools like MADGRAPH [31], CALCHEP [32] / COMPHEP [33], and several others [34–40]. Here, we require only that the reader has a basic familiarity with the methods involved from graduate-level particle physics courses based, for example, on Refs. [1, 3]. Our main concerns are the uses to which these calculations are put, their limitations and ways to improve on the results obtained with them.

[11] In the context of hadronization models, this coherent superposition of two quarks in an overall antitriplet state is sometimes called a 'diquark' (at low m_{qq}) or a 'string junction' (at high m_{qq}), see Section 5.1; it corresponds to the antisymmetric '$\mathrm{red} + \mathrm{green} = \mathrm{antiblue}$' combination needed to create a baryon wavefunction.

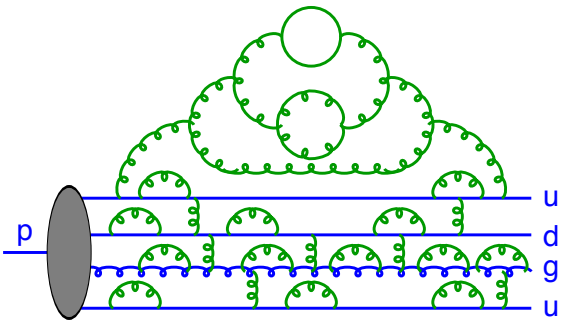


Fig. 8: Illustration of partonic fluctuations inside a proton beam (from Ref. [41])

For illustration, take one of the most commonly occurring processes in hadron collisions: Rutherford scattering of two quarks via a t-channel gluon exchange—Fig. 7—which at leading order has the differential cross section

$$\mathrm{qq}' \to \mathrm{qq}' \quad : \quad \frac{\mathrm{d}\sigma}{\mathrm{d}\hat{t}} = \frac{\pi}{\hat{s}^2}\,\frac{4}{9}\,\alpha_\mathrm{s}^2\,\frac{\hat{s}^2+\hat{u}^2}{\hat{t}^2}, \tag{36}$$

with the $2 \to 2$ Mandelstam variables ('hatted' to emphasize that they refer to a partonic $2 \to 2$ scattering rather than the full $\mathrm{pp} \to$ jets process)

$$\hat{s} = (p_1+p_2)^2\,, \tag{37}$$

$$\hat{t} = (p_3-p_1)^2 = -\hat{s}\frac{(1-\cos\hat{\theta})}{2}, \tag{38}$$

$$\hat{u} = (p_4-p_1)^2 = -\hat{s}\frac{(1+\cos\hat{\theta})}{2}. \tag{39}$$

Reality, however, is more complicated; the picture on the right-hand pane of Fig. 7 shows a real dijet event, as recorded by the ATLAS experiment. The complications to be addressed when going from left to right in Fig. 7 are: (1) additional jets, a.k.a. real-emission corrections, which can significantly change the topology of the final state, potentially shifting jets in or out of an experimentally defined acceptance region; (2) loop factors, a.k.a. virtual corrections, change the number of available quantum paths through phase space, and hence modify the normalization of the cross section (total *and* differential); and (3) additional corrections are generated by confinement and by the so-called underlying event. These corrections must be taken into account to complete our understanding of QCD and connect the short-distance physics with macroscopic experiments. Apart from the perturbative expansion itself, the most powerful tool we have to organize this vast calculation, is factorization.

2.1 Factorization

In high-energy scattering problems involving hadrons in the initial state, we immediately face the complication that hadrons are composite, with a time-dependent structure illustrated in Fig. 8; there are partons within clouds of further partons, constantly being emitted and absorbed. Thus, before we can use perturbatively calculated partonic-scattering MEs, we must first address the partonic structure of the colliding hadron(s).

For the hadron to remain intact, the fluctuations inside it must involve momentum transfers smaller than the confinement scale. Indeed, high-virtuality fluctuations are suppressed by powers of

$$\frac{\alpha_\mathrm{s}\,\Lambda^2}{|k|^2}, \tag{40}$$

with Λ the confinement scale (~ 200 MeV, see Section 1.4) and $|k|$ the virtuality of the fluctuation. Thus, most fluctuations occur over timescales of $\sim 1/\Lambda$.

A hard perturbative probe, on the other hand, such as the exchanged photon in DIS (Fig. 3), interacts over a much shorter timescale $1/Q \ll 1/\Lambda$, during which the partonic fluctuations in the struck hadron appear almost frozen. The hard probe effectively takes an instantaneous snapshot of the hadron structure, at a characteristic resolution given by $\sim 1/Q$.

This is formalized by the *factorization theorem* [42] (see also the TASI lectures by George Sterman [43]), which expresses the independence of long-wavelength (soft) structure on the nature of the hard (short-distance) process. Originally formulated for DIS, factorization allows us to write the cross section for lepton–hadron scattering as a convolution of a non-perturbative but universal (i.e., process-independent) parton density function (PDF) and a perturbatively calculable partonic-scattering cross section. Denoting the fraction of the hadron momentum carried by parton i by x_i,

$$\vec{p}_i \;=\; x_i\,\vec{p}_h, \tag{41}$$

we may write the lepton–hadron cross section in factorized form (see, e.g., Refs. [3, 44]),

$$\sigma_{\ell h} = \sum_i \int_0^1 \mathrm{d}x_i \int \mathrm{d}\Phi_f \, f_{i/h}(x_i,\mu_F^2)\,\frac{\mathrm{d}\hat{\sigma}_{\ell i\to f}(x_i,\Phi_f,\mu_F^2)}{\mathrm{d}x_i\,\mathrm{d}\Phi_f}, \tag{42}$$

with i an index running over all possible parton types[12] in the incoming hadron and f enumerating all possible (partonic) final states, with Lorentz-invariant phase space, Φ_f.

The *PDFs*, $f_{i/h}$, parameterize the distribution of partons inside the target hadron. They are not a priori calculable and must be constrained by fits to data. This is discussed in Section 2.2.

The *partonic cross section*, $\mathrm{d}\hat{\sigma}$, knows nothing of the target hadron apart from the fact that it contained the struck parton. It is calculable within perturbation theory, as will be discussed in Section 2.3.

The dividing line between the two is drawn at an arbitrary ('user-defined') scale, μ_F, called the *factorization scale*. There is some arbitrariness involved in choosing a value for μ_F. Some heuristic arguments to guide in the choice of factorization scale are the following. On the long-distance side, the PDFs include a (re)summation of fluctuations inside fluctuations up to virtualities of order μ_F. It would, therefore, not make much sense to take μ_F significantly larger than the scales characterizing resolved particles on the short-distance side of the calculation (i.e., the particles appearing explicitly in Φ_f); otherwise the PDFs would be including sums over fluctuations that happen on timescales shorter than those probed by the physical process. Similarly, μ_F should also not be taken much lower than the scale(s) appearing in the hard process. For MEs characterized by a single well-defined scale, such as the Q^2 scale in DIS or the invariant-mass scale $\hat{s}$ in Drell–Yan production ($\mathrm{q\bar{q}} \to \mathrm{Z}/\gamma^* \to \ell^+\ell^-$), such arguments essentially fix the preferred scale choice to $\mu_F = Q$ or $\mu_F = \sqrt{\hat{s}}$, respectively, which may then be varied by a factor of two (or larger) around the nominal value in order to estimate uncertainties. For multi-scale problems, however, such as $\mathrm{pp} \to \mathrm{Z/W} + \mathrm{n\,jets}$, there are several a priori equally good choices available, from the lowest to the highest QCD scales that can be constructed from the final-state momenta, usually with several dissenting groups of theorists arguing over which particular choice is best. Suggesting that one might simply *measure* the scale would not really be an improvement, as the factorization scale is fundamentally unphysical and therefore unobservable (similarly to gauge or convention choices). One plausible strategy is to look at higher-order (NLO or NNLO) calculations, in which correction terms appear that cancel the dependence on the scale choice, stabilizing the final result. From such comparisons, a 'most stable' initial scale choice can, in principle, be determined, which then furnishes a reasonable starting point, but we emphasize that the question *is* intrinsically ambiguous, and

[12] Typically, only quarks and gluons are included in this sum, but also photons and even leptons can, in principle, be included. Similarly, PDFs are normally used to describe hadrons, but can also be defined, for example, to describe the cloud of virtual photons (and fermion pairs) surrounding an electron.

no golden recipe is likely to magically give all the right answers. The best we can do is to vary the value of μ_F not only by an overall factor, but also by exploring different possible forms for its functional dependence on the momenta appearing in Φ_f. A complementary useful discussion of the pros and cons of different factorization scale choices can be found in the TASI lectures by Tilman Plehn [45].

Secondly, and more technically, at NLO and beyond one also has to settle on a *factorization scheme* in which to do the calculations. For all practical purposes, students focusing on LHC physics are only likely to encounter one such scheme, the modified minimal-subtraction ($\overline{\rm MS}$) scheme already mentioned in the discussion of the definition of the strong coupling in Section 1.4. At the level of these lectures, we shall therefore not elaborate further on this choice here.

We note that factorization can also be applied multiple times, to break up a complicated calculation into simpler pieces that can be treated as approximately independent. This will be very useful when dealing with successive emissions in a parton shower, Section 3.2, or when factoring off decays of long-lived particles from a hard production process, Section 3.4.

We round off the discussion of factorization by mentioning a few caveats the reader should be aware of. (See Ref. [43] for a more technical treatment.)

Firstly, the proof only applies to the first term in an operator product expansion in 'twist' $=$ $\text{mass dimension} - \text{spin}$. Since operators with higher mass dimensions are suppressed by the hard scale to some power, this leading twist approximation becomes exact in the limit $Q \to \infty$, while at finite Q it neglects corrections of order

$$\text{Higher Twist}: \frac{[\ln(Q^2/\Lambda^2)]^{m<2n}}{Q^{2n}} \quad (n = 2 \text{ for DIS}). \tag{43}$$

In Section 5, we shall discuss some corrections that go beyond this approximation, in the context of multiple parton–parton interactions.

Secondly, the proof only really applies to inclusive cross sections in DIS [42] and in Drell–Yan [46]. For all other hadron-initiated processes, factorization is an ansatz. For a general hadron–hadron process, we write the assumed factorizable cross section as

$$\mathrm{d}\sigma_{h_1 h_2} \;=\; \sum_{i,j} \int_0^1 \mathrm{d}x_i \int_0^1 \mathrm{d}x_j \sum_f \int \mathrm{d}\Phi_f \, f_{i/h_1}(x_i, \mu_F^2) \, f_{j/h_2}(x_j, \mu_F^2) \, \frac{\mathrm{d}\hat{\sigma}_{ij\to f}}{\mathrm{d}x_i \, \mathrm{d}x_j \, \mathrm{d}\Phi_f}. \tag{44}$$

Note that, if $\mathrm{d}\hat{\sigma}$ is divergent (as, e.g., Rutherford scattering is) then the integral over $\mathrm{d}\Phi_f$ must be regulated, for example, by imposing some explicit minimal transverse-momentum cut and/or other phase-space restrictions.

2.2 Parton densities

The PDF, $f_{i/h}(x,\mu_F^2)$, represents the effective density of partons of type/flavour i, as a function of the momentum fraction[13] x_i, when a hadron of type h is probed at the factorization scale μ_F. The PDFs are non-perturbative functions which are not a priori calculable, but a perturbative differential equation governing their evolution with μ_F can be obtained by requiring that physical-scattering cross sections, such as the one for DIS in Eq. (42), be independent of μ_F to the calculated orders [47]. The resulting *renormalization group equation* (RGE) is called the DGLAP[14] equation and can be used to 'run' the PDFs from one perturbative resolution scale to another (its evolution kernels are the same as those used in parton showers, to which we return in Section 3.2).

This means that we only need to determine the form of the PDF as a function of x in a single (arbitrary) scale, μ_0. We can then get its form at any other scale μ_F by simple RGE evolution. In the

[13] Recall: the x fraction is defined in Eq. (41).

[14] DGLAP: Dokshitzer–Gribov–Lipatov–Altarelli–Parisi [47–49].

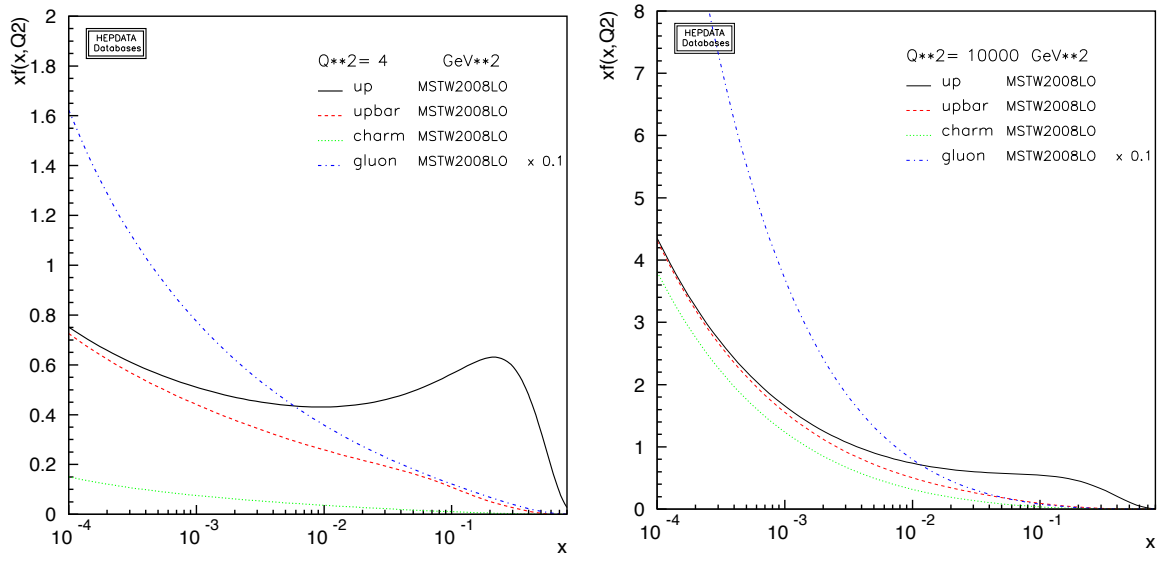


Fig. 9: Illustration of the change of the u (black), $\bar{u}$ (red, dashed), c (green, dotted), and g (blue, dot-dashed) distributions, from $Q = \mu_F = 2\,\text{GeV}$ (left) to $Q = \mu_F = 100\,\text{GeV}$ (right). Note that a factor 0.1 has been applied to the gluon distribution. Plots made using the HEPDATA online tool [55].

context of PDF fits (constraining the form of the PDF functions by fitting cross sections to experimental data, for example, from DIS [50, 51], Drell–Yan [52, 53], and $\text{pp} \to \text{jets}$ [54]), the reference scale μ_0 is usually taken to be relatively low, of order one or a few GeV.

The behaviour of the PDFs as we evolve μ_F from a low scale, 2 GeV, to a high one, 100 GeV, is illustrated in Fig. 9, for the MSTW[15] 2008 LO[16] PDF set [56]. At low $Q = \mu_F = 2\,\text{GeV}$ (left), the proton structure is dominated by a few hard quarks (a 'valence bump' is clearly visible around $x \sim 0.2$), while at higher scales $Q = 100\,\text{GeV}$ (right) we predominantly resolve fluctuations within fluctuations, yielding increasingly large gluon- and sea-quark distributions with rather small x values, while the valence quarks play a progressively smaller role.

We note that different collaborations, like CTEQ, MSTW, NNPDF, etc., use different ansätze for the form of $f(x, \mu_0^2)$. They may also include different data in the fits, and/or treat or weight the data differently. Thus, results from different groups may not always be mutually compatible. An example is given in Fig. 10, which shows the difference between the CTEQ6L1 gluon PDF [57] (red dashed) and the MSTW 2008 LO PDF [56], normalized to MSTW (which would thus be a flat line at zero), at $\mu_F = 10\,\text{GeV}$. The y-axis shows the relative difference between the sets, in percent. Also shown are the 90% CL contours computed from the uncertainty variations included in the MSTW 2008 LO set (black). Using only the MSTW uncertainty band, one would arrive at an estimated $\sim 5\%$ uncertainty over most of the x range, while including the CTEQ6L1 set would increase that to $\sim 10\%$. At NLO, this discrepancy is reduced, but not removed. A significant effort is currently being undertaken within the PDF community to agree on common, and more comprehensive, ways of defining PDF uncertainty bands [54, 58]. This is complicated due to the different ways of defining $f(x, \mu_0^2)$ and due to the experimental data sets not always being fully compatible with one another. For the time being, it is recommended to try at least sets from two different groups, for a comprehensive uncertainty estimate.

Occasionally, the words *structure functions* and *parton densities* are used interchangeably. However, there is an important distinction between the two, which we find often in (quantum) physics: the

[15] MSTW: Martin–Stirling–Thorne–Watt.

[16] The 'LO' means that the fit was performed using LO MEs in the cross section formulae.

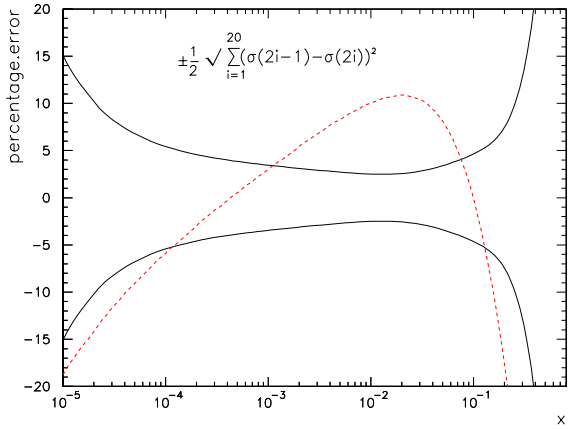


Fig. 10: Illustration of the difference between the MSTW 2008 and CTEQ6 LO gluon PDFs at $\mu_F = 10\,\mathrm{GeV}$. All curves are normalized to the central MSTW 2008 prediction. The black solid lines show the 90% CL MSTW variations, while the dashed red line shows the CTEQ6L1 distribution.

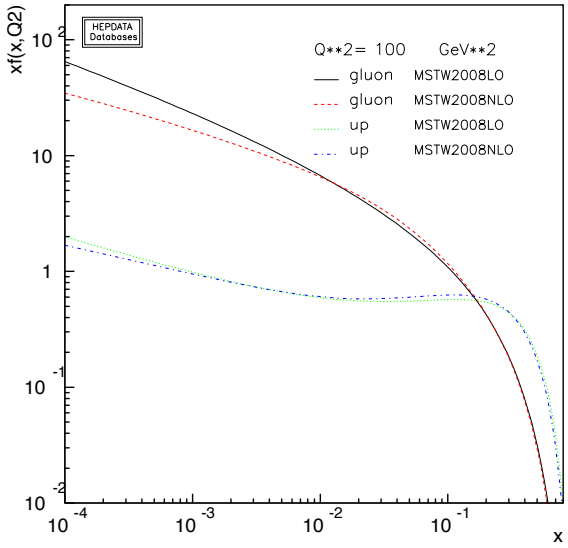


Fig. 11: Illustration of the change between PDF fits using LO and NLO MEs: the g distribution at LO (black) and NLO (red, dashed), and the u distribution at LO (green, dotted) and NLO (blue, dot-dashed), for the MSTW 2008 PDF sets [56], at $Q = \mu_F = 10\,\mathrm{GeV}$. Plots made using the HEPDATA online tool [55].

former is a physical observable used to parameterize the DIS cross sections (see, e.g., Ref. [3]), while the latter is a 'fundamental' quantity extracted from it. In particular, since the parton densities are not, themselves, physically observable, they can only be defined within a specific factorization scheme, order by order in perturbation theory. The only exception is at leading order, at which they have the simple physical interpretation of parton-number densities. When going to higher orders, we tend to keep the simple intuitive picture from LO in mind, but one should be aware that the fundamental relationship between PDFs and measured quantities is now more complicated (due to the interplay between the PDFs and the real and virtual corrections to the LO cross section), and that the parton densities no longer have a clear probabilistic interpretation starting from NLO.

The reader should also be aware that there is some ambiguity whether NLO PDFs should be used for LO calculations. In principle, the higher-order PDFs are better constrained and the difference

between, for example, an NLO and an LO set should formally be beyond LO precision, so that one might be tempted to simply use the highest-order available PDFs for any calculation. However, higher-order terms can sometimes be absorbed, at least partially, into effective lower-order coefficients. In the context of PDFs, the fit parameters of lower-order PDFs will attempt to compensate for missing higher-order contributions in the matrix elements. To the extent those higher-order contributions are *universal*, this is both desirable and self-consistent. This leads to some typical qualitative differences between LO and NLO PDFs, illustrated in Fig. 11: NLO PDFs tend to be smaller at low x and slightly larger at high x, than LO ones. Thus, it is quite possible that using an NLO PDF in conjunction with LO MEs can give a worse agreement with data than LO PDFs do.

Finally, another oft-raised question concerns which PDF sets to use for the parton-shower evolution in Monte Carlo generators. Importantly, the equations driving the initial-state showers in Monte Carlo models are only sensitive to *ratios* of PDFs [59]. Since the shower evolution typically only has leading-logarithmic (LL) precision, it should be theoretically consistent to use any (LO or better) PDF set to drive the evolution. However, similarly to above, there will be subleading differences between different choices, and one is justified in worrying about the level of physical effects that could be generated. Unfortunately, there is currently no way to ensure 100% self-consistency. Since PDF fits are not done with MC codes, but instead use analytical resummation models (see, e.g., the TASI lectures by Sterman [43]), which are not identical to their MC counterparts, the PDF fits are essentially 'tuned' to a slightly different resummation than that incorporated in a given MC model. In practice, not much is known about the size and impact of this ambiguity [60]. Known differences include: the size of phase space (purely collinear massless PDF evolution versus the finite-transverse-momentum massive MC phase space), the treatment of momentum conservation and recoil effects, additional higher-order effects explicitly or implicitly included in the MC evolution, choice of renormalization scheme and scale, and, for those MC algorithms that do not rely on collinear (DGLAP, see Ref. [3]) splitting kernels (e.g., the various kinds of dipole evolution algorithms, see Ref. [61]), differences in the effective factorization scheme.

As a baseline, we recommend simply using whatever PDF set the given MC model was originally tuned with, since this should de facto (by fitting the available data) reabsorb as much of the inconsistency as possible. Furthermore, it should be emphasized that underlying-event and minimum-bias models based on multi-parton interactions (see Section 5.2) usually make the explicit assumption that the PDFs can be interpreted as physical number densities even down to very low Q and x, a property which is generally only true for LO PDFs. It must therefore be strongly discouraged to use (N)NLO PDF sets in this context.

2.3 Fixed-order QCD

Consider the production of an arbitrary final state, F (e.g., a Higgs boson, a $\mathrm{t\bar{t}}$ pair, etc). Schematically, we may express the (perturbative) all-orders differential cross section for an observable $\mathcal{O}$, in the following way:

$$\left.\frac{\mathrm{d}\sigma_F}{\mathrm{d}\mathcal{O}}\right|_{\mathrm{ME}} = \underbrace{\sum_{k=0}^{\infty}\int \mathrm{d}\Phi_{F+k}}_{\Sigma\text{ legs}} \underbrace{\left|\sum_{\ell=0}^{\infty}\mathcal{M}_{F+k}^{(\ell)}\right|^2}_{\Sigma\text{ loops}} \delta\left(\mathcal{O}-\mathcal{O}(\Phi_{F+k})\right), \tag{45}$$

where, for compactness, we have suppressed all PDF and luminosity normalization factors. $\mathcal{M}_{F+k}^{(\ell)}$ is the amplitude for producing F in association with k additional final-state partons, 'legs' and with ℓ additional loops. The sums start at $k=0$ and $\ell=0$, corresponding to the leading order for producing F, while higher terms represent real and virtual corrections, respectively.

The purpose of the δ function is to project out hypersurfaces of constant value of $\mathcal{O}$ in the full $\mathrm{d}\Phi_{F+k}$ phase space, with $\mathcal{O}(\Phi_{F+k})$ a function that defines $\mathcal{O}$ evaluated on each specific momentum

configuration, Φ_{F+k}. (Without the δ function, the formula would give the total integrated cross section, instead of the cross section differentially in $\mathcal{O}$.)

We recover the various fixed-order truncations of perturbative QCD (pQCD) by limiting the nested sums in Eq. (45) to include only specific values of $k + \ell$. Thus,

$$
\begin{aligned}
k = 0, \ell = 0 \quad &\Longrightarrow \quad \text{leading order (usually tree-level) for } F \text{ production} \\
k = n, \ell = 0 \quad &\Longrightarrow \quad \text{leading order for } F + n \text{ jets} \\
k + \ell \leq n, \quad &\Longrightarrow \quad \text{N}^n\text{LO for } F \text{ (includes N}^{n-1}\text{LO for } F + 1 \text{ jet, N}^{n-2}\text{LO for } F + \\
&\qquad\qquad 2 \text{ jets, and so on up to LO for } F + n \text{ jets)} .
\end{aligned}
$$

For $k \geq 1$, we are not considering inclusive F production; instead, we are considering the process $F + k$ jets. If we simply integrate over all momenta, as implied by the integration over $\mathrm{d}\Phi_{F+k}$ in Eq. (45), we would be including configurations in which one or more of the k partons are collinear or soft. Such configurations are IR divergent in QCD and hence must be regulated. Since we talk about *collinear* and *soft* divergences (the origins of which will be discussed in more detail in Sections 2.4 and 3.2), cuts on *angles* and *energies* and/or cuts on combinations, like *transverse momenta*, can be used to cut away the problematic regions of phase space.

Recall, however, that pQCD is approximately scale invariant. This implies that a regularization cut on a dimensionful quantity, like energy or transverse momentum, should be formulated as a *ratio* of scales, rather than as an absolute number. For example, a jet with $p_\perp = 50\,\mathrm{GeV}$ would be considered hard and well-separated if produced in association with an ordinary Z boson (with hard scale $M_\mathrm{Z} = 91.2\,\mathrm{GeV}$), while the same jet would be considered soft if produced in association with a 900 GeV Z′ boson (see Refs. [18, 19] for more explicit examples).

The essence of the point is that, if the regularization scale is taken too low, logarithmic enhancements of the type

$$
\alpha_\mathrm{s}^n \ln^{m \leq 2n}\left(\frac{Q_F^2}{Q_k^2}\right) \tag{46}
$$

will generate progressively larger corrections, order by order, which will spoil any fixed-order truncation of the perturbative series. Here, Q_F is the hard scale associated with the process under consideration, while Q_k is the scale associated with an additional parton, k.

A good rule of thumb is that if $\sigma_{k+1} \approx \sigma_k$ (at whatever order you are calculating), then the perturbative series is converging too slowly for a fixed-order truncation of it to be reliable. For fixed-order perturbation theory to be applicable, you must place your cuts on the hard process such that $\sigma_{k+1} \ll \sigma_k$. In the discussion of parton showers in Section 3.2, we shall see how the region of applicability of perturbation theory can be extended.

The virtual amplitudes, for $\ell \geq 1$, are divergent for any point in phase space. However, as encapsulated by the famous KLN theorem [62,63], unitarity (which essentially expresses probability conservation) puts a powerful constraint on the infrared (IR) divergences[17], forcing them to cancel exactly against those coming from the unresolved real emissions that we had to cut out above, order by order, making the complete answer for fixed $k + \ell = n$ finite[18] Nonetheless, since this cancellation happens between contributions that formally live in different phase spaces, a main aspect of loop-level higher-order calculations is how to arrange for this cancellation in practice, either analytically or numerically, with many different methods currently on the market. We shall discuss the idea behind subtraction approaches in Section 2.4.

[17] The loop integrals also exhibit ultra-violet (UV) divergences, but these are dealt with by renormalization.

[18] Formally, the KLN theorem states that the sum over degenerate quantum states is finite. In the context of fixed-order perturbation theory, this is exemplified by states with infinitely collinear and/or soft radiation being degenerate with the corresponding states with loop corrections; they cannot be told apart by any physical observable.

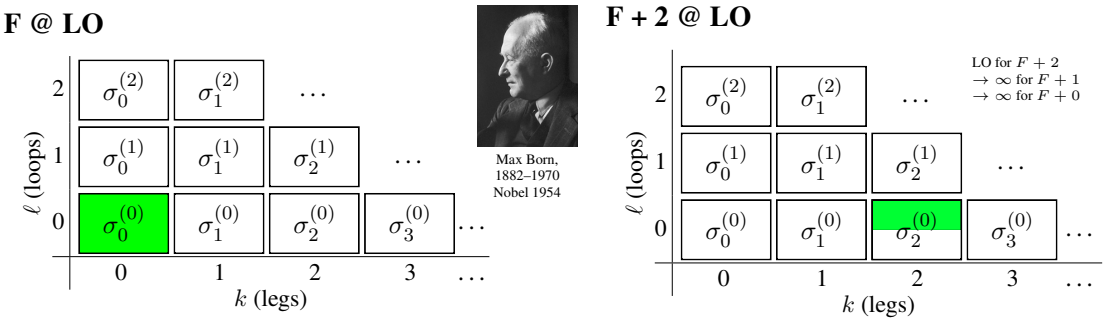


Fig. 12: Coefficients of the perturbative series covered by LO calculations. *Left:* F production at LO. *Right:* $F + 2$ jets at LO, with the half-shaded box illustrating the restriction to the region of phase space with exactly 2 resolved jets. The total power of α_s for each coefficient is $n = k+\ell$. (Photo of Max Born from `nobelprize.org`.)

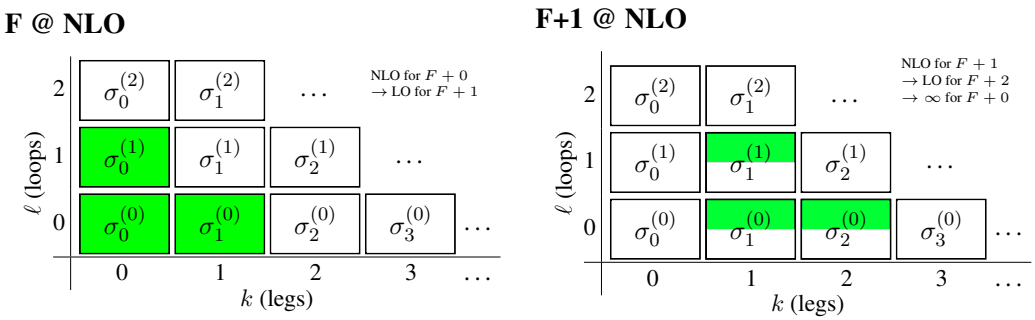


Fig. 13: Coefficients of the perturbative series covered by NLO calculations. *Left:* F production at NLO. *Right:* $F + 1$ jet at NLO, with half-shaded boxes illustrating the restriction to the region of phase space with exactly one resolved jet. The total power of α_s for each coefficient is $n = k + \ell$.

A convenient way of illustrating the terms of the perturbative series that a given ME-based calculation includes is given in Fig. 12. In the left-hand pane, the shaded box corresponds to the lowest-order 'Born-level' ME squared. This coefficient is non-singular and hence can be integrated over all of phase space, which we illustrate by letting the shaded area fill all of the relevant box. A different kind of leading-order calculation is illustrated in the right-hand pane of Fig. 12, where the shaded box corresponds to the LO ME squared for $F + 2$ jets. This coefficient diverges in the part of phase space where one or both of the jets are unresolved (i.e., soft or collinear), and hence integrations can only cover the hard part of phase space, which we reflect by only shading the upper half of the relevant box.

Figure 13 illustrates the inclusion of NLO virtual corrections. To prevent confusion, first a point on notation: by $\sigma_0^{(1)}$, we intend

$$\sigma_0^{(1)} = \int \mathrm{d}\Phi_0 \; 2\mathrm{Re}[\mathcal{M}_0^{(1)}\mathcal{M}_0^{(0)*}], \tag{47}$$

which is of order α_s relative to the Born level. Compare, for example, with the expansion of Eq. (45) to order $k+\ell = 1$. In particular, $\sigma_0^{(1)}$ should *not* be confused with the integral over the one-loop ME squared (which would be of relative order α_s^2 and hence forms part of the NNLO coefficient $\sigma_0^{(2)}$). Returning to Fig. 13, the unitary cancellations between real and virtual singularities imply that we can now extend the integration of the real correction in the left-hand pane over all of phase space, while retaining a finite

F @ NNLO

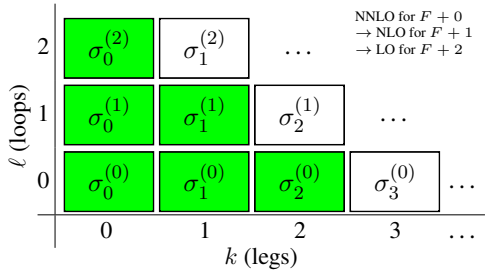


Fig. 14: Coefficients of the perturbative series covered by an NNLO calculation. The total power of α_s for each coefficient is $n = k + \ell$. Green shading represents the full perturbative coefficient at the respective k and ℓ.

total cross section,

$$\begin{aligned} \sigma_0^{\mathrm{NLO}} &= \int \mathrm{d}\Phi_0\, |\mathcal{M}_0^{(0)}|^2 + \int \mathrm{d}\Phi_1\, |\mathcal{M}_1^{(0)}|^2 + \int \mathrm{d}\Phi_0\; 2\mathrm{Re}[\mathcal{M}_0^{(1)}\mathcal{M}_0^{(0)*}] \\ &= \sigma_0^{(0)} + \sigma_1^{(0)} + \sigma_0^{(1)}, \end{aligned} \tag{48}$$

with $\sigma_0^{(0)}$ the finite Born-level cross section, and the positive divergence caused by integrating the second term over all of phase space is cancelled by a negative divergence coming from the integration over loop momenta in the third term. One method for arranging the cancellation of singularities—subtraction—is discussed in Section 2.4.

However, if our starting point for the NLO calculation is a process which already has a non-zero number of hard jets, we must continue to impose that at least that number of jets must still be resolved in the final-state integrations,

$$\begin{aligned} \sigma_1^{\mathrm{NLO}}(p_{\perp\mathrm{min}}) &= \int_{p_\perp > p_{\perp\mathrm{min}}} \mathrm{d}\Phi_1\, |\mathcal{M}_1^{(0)}|^2 + \int_{p_{\perp 1} > p_{\perp\mathrm{min}}} \mathrm{d}\Phi_2\, |\mathcal{M}_2^{(0)}|^2 + \int_{p_\perp > p_{\perp\mathrm{min}}} \mathrm{d}\Phi_1\; 2\mathrm{Re}[\mathcal{M}_1^{(1)}\mathcal{M}_1^{(0)*}] \\ &= \sigma_1^{(0)}(p_\perp > p_{\perp\mathrm{min}}) + \sigma_2^{(0)}(p_{\perp 1} > p_{\perp\mathrm{min}}) + \sigma_1^{(1)}(p_\perp > p_{\perp\mathrm{min}}), \end{aligned} \tag{49}$$

where the restriction to at least one jet having $p_\perp > p_{\perp\mathrm{min}}$ has been illustrated in the right-hand pane of Fig. 13 by shading only the upper part of the relevant boxes. In the second term in Eq. (49), the notation $p_{\perp 1}$ is used to denote that the integral runs over the phase space in which at least one 'jet' (which may consist of one or two partons) must be resolved with respect to $p_{\perp\mathrm{min}}$. Here, therefore, an explicit dependence on the algorithm used to define 'a jet' enters for the first time. This is discussed in more detail in the 2009 ESHEP lectures by Salam [4].

To extend the integration to cover also the case of two unresolved jets, we must combine the left- and right-hand parts of Fig. 13 and add the new coefficient

$$\sigma_0^{(2)} = |\mathcal{M}_0^{(1)}|^2 + 2\mathrm{Re}[\mathcal{M}_0^{(2)}\mathcal{M}_0^{(0)*}], \tag{50}$$

as illustrated by the diagram in Fig. 14.

2.4 The subtraction idea

According to the KLN theorem, the IR singularities coming from integrating over collinear and soft real-emission configurations should cancel, order by order, by those coming from the IR divergent-loop integrals. This implies that we should be able to rewrite, for example, the NLO cross section, Eq. (48), as

$$\sigma^{\mathrm{NLO}} = \sigma^{\mathrm{Born}} + \mathrm{Finite}\left\{\int \mathrm{d}\Phi_{F+1}\, |\mathcal{M}_{F+1}^{(0)}|^2\right\} + \mathrm{Finite}\left\{\int \mathrm{d}\Phi_F\; 2\mathrm{Re}[\mathcal{M}_F^{(1)}\mathcal{M}_F^{(0)*}]\right\}, \tag{51}$$

with the second and third terms having had their common (but opposite-sign) singularities cancelled out and some explicitly finite quantities remaining.

The first step towards this goal is to classify all IR singularities that could appear in the amplitudes. We know that the IR limits are universal, so they can be classified using a set of process-independent functions that only has to be worked out once and for all. A widely used such set of functions are the *Catani–Seymour* (CS) dipole functions [64, 65], a method which by now has even been partially automated [66, 67]. Here, we shall instead use a formalism based on *antennae* [68–70]. The distinction between the two is basically that one antenna is made up of two dipole 'ends', and hence the antenna formalism tends to generate somewhat fewer terms. At NLO, however, there is no fundamental incompatibility—the antennae we use here can always be partitioned into two dipole ends, if so desired. (Note: only the antenna method has been successfully generalized to NNLO [71, 72]. Other NNLO techniques, not covered here, are *sector decomposition*, see Refs. [73, 74], and the generic formalism for hadroproduction of colourless states presented in Ref. [75].)

At NLO, the idea with subtraction is thus to rewrite the NLO cross section by adding and subtracting a simple function, $\mathrm{d}\sigma_S$, that encapsulates all the IR limits,

$$\sigma^{\mathrm{NLO}} = \sigma^{\mathrm{Born}} + \int \mathrm{d}\Phi_{F+1} \underbrace{\left(|\mathcal{M}^{(0)}_{F+1}|^2 - \mathrm{d}\sigma^{\mathrm{NLO}}_S\right)}_{\text{finite by universality}} + \underbrace{\int \mathrm{d}\Phi_F \; 2\mathrm{Re}[\mathcal{M}^{(1)}_F \mathcal{M}^{(0)*}_F] + \int \mathrm{d}\Phi_{F+1} \, \mathrm{d}\sigma^{\mathrm{NLO}}_S}_{\text{finite by KLN}} . \tag{52}$$

The task now is to construct a suitable form for $\mathrm{d}\sigma_S$. A main requirement is that it should be sufficiently simple that the integral in the last term can be done analytically, in dimensional regularization, so that the IR poles it generates can be cancelled against those from the loop term.

To build a set of universal terms that parameterize the IR singularities of any amplitude, we start from the observation that gauge theory amplitudes factorize in the *soft limit*, as follows:

$$|\mathcal{M}_{F+1}(\ldots,i,j,k,\ldots)|^2 \overset{j_g \to 0}{\to} g_s^2 \, N_{\mathrm{C}} \left(\frac{2 s_{ik}}{s_{ij} s_{jk}} - \frac{2m_i^2}{s_{ij}^2} - \frac{2m_k^2}{s_{jk}^2} \right) |\mathcal{M}_F(\ldots,i,k,\ldots)|^2 \ , \tag{53}$$

where parton j is a soft gluon, partons i, j, and k form a chain of colour-space index contractions (we say they are *colour-connected*), g_{s} is the strong coupling, and the terms in parentheses are called the *soft-eikonal factor*. We here show it including mass corrections, which appear if i and k have non-zero rest masses, with the invariants s_{ab} then defined as

$$s_{ab} \equiv 2 p_a \cdot p_b = (p_a + p_b)^2 - m_a^2 - m_b^2 \, . \tag{54}$$

The colour factor, N_{C}, is valid for the leading-colour contribution, regardless of whether the i and k partons are quarks or gluons. At subleading colour, an additional soft-eikonal factor identical to the one above but with a colour factor proportional to $-1/N_{\mathrm{C}}$ arises for each $\mathrm{q\bar{q}}$ pair combination. This, for example, modifies the effective colour factor for $\mathrm{q\bar{q}} \to \mathrm{qg\bar{q}}$ from N_{C} to $N_{\mathrm{C}}(1 - 1/N_{\mathrm{C}}) = 2C_F$, in agreement with the colour factor for quarks being C_F rather than C_A.

Similarly, amplitudes also factorize in the *collinear limit* (partons i and j are parallel, so $s_{ij} \to 0$), in which the eikonal factor above is replaced by the famous DGLAP splitting kernels [47–49], which were already mentioned in Section 2.2, in the context of PDF evolution. They are also the basis of conventional parton-shower models, to which we return in Section 3.2.

Essentially, what antenna functions, CS dipoles, and the like, all do, is to combine the soft (eikonal) and collinear (Altarelli–Parisi) limits into one universal set of functions that achieve the correct limiting

behaviour for *both* soft and collinear radiation. To give an explicit example, the *antenna function* for gluon emission from a colour-connected $\mathrm{q\bar{q}}$ pair can be derived from the MEs squared for the process $\mathrm{Z}^0 \to \mathrm{q\bar{q}} \to \mathrm{qg\bar{q}}$ [76],

$$\frac{\left|\mathcal{M}(Z^0 \to q_i g_j \bar{q}_k)\right|^2}{\left|\mathcal{M}(Z^0 \to q_I \bar{q}_K)\right|^2} \;=\; g_s^2\, 2C_F \left[\underbrace{\frac{2s_{ik}}{s_{ij}s_{jk}}}_{\text{eikonal}} + \frac{1}{s_{IK}} \underbrace{\left(\frac{s_{jk}}{s_{ij}} + \frac{s_{ij}}{s_{jk}} \right)}_{\text{collinear}} \right], \tag{55}$$

where we have neglected mass corrections (see Refs. [77, 78] for massive expressions) and we recognize the universal eikonal soft factor from Eq. (53) in the first term. The two additional terms are less singular, and are required to obtain the correct collinear (Altarelli–Parisi) limits as $s_{ij} \to 0$ or $s_{jk} \to 0$.

However, since the singularity structure is universal, we could equally well have used the process $\mathrm{H}^0 \to \mathrm{q\bar{q}} \to \mathrm{qg\bar{q}}$ to derive the antenna function. Our antenna function would then have come out as [78],

$$\frac{\left|\mathcal{M}(\mathrm{H}^0 \to \mathrm{q}_i \mathrm{g}_j \bar{\mathrm{q}}_k)\right|^2}{\left|\mathcal{M}(\mathrm{H}^0 \to \mathrm{q}_I \bar{\mathrm{q}}_K)\right|^2} \;=\; g_s^2\, 2C_F \left[\underbrace{\frac{2s_{ik}}{s_{ij}s_{jk}}}_{\text{eikonal}} + \frac{1}{s_{IK}} \underbrace{\left(\frac{s_{jk}}{s_{ij}} + \frac{s_{ij}}{s_{jk}} \right)}_{\text{collinear}} + \underbrace{\frac{2}{s_{IK}}}_{\text{finite}} \right], \tag{56}$$

where the additional term, $2/s_{IK}$, is non-singular ('finite') over all of phase space. Thus, we here see an explicit example that the singularities are process independent while the non-singular terms are process dependent. Since we add and subtract the same term in Eq. (52), the final answer does not depend on the choice of finite terms. We say that they correspond to different *subtraction schemes*. One standard antenna subtraction scheme, which uses the antenna function defined in Eq. (55) rather than the one in Eq. (56), is the Gehrmann–Gehrmann–de Ridder–Glover (GGG) one, given in Ref. [70].

If there is more than one colour antenna in the Born-level process, the form of $\mathrm{d}\sigma_S$ is obtained as a sum over terms, each of which captures one specific soft limit and either all or 'half' of a collinear one, depending on the specific scheme and the type of parton,

$$\mathrm{d}\sigma_S \;=\; \sum_j A_{IK\to ijk}\, \left|\mathcal{M}_F(\ldots, I, K, \ldots)\right|^2, \tag{57}$$

with the sum running over all singular $3 \to 2$ 'clusterings' of the $(F+1)$-parton state to F partons. An analysis of the different ways of partitioning the collinear singularity of gluons among neighbouring antenna is beyond the scope of this introduction, but useful discussions can be found in Ref. [79–81].

2.5 Infrared safety

A further requirement for being able to perform calculations within pQCD is that the observable be *IR safe*. Note that by 'IR', we here mean any limit that involves a low scale (i.e., any non-UV limit), without regard to whether it is collinear or soft.

The property of IR safety defines a special class of observables which have *minimal sensitivity* to long-distance physics, and which can be consistently computed in pQCD. An observable is IR safe if:

1. *(safety against soft radiation):* adding any number of infinitely soft particles should not change the value of the observable;
2. *(safety against collinear radiation):* splitting an existing particle up into two co-moving particles, with arbitrary fractions z and $1-z$, respectively, of the original momentum, should not change the value of the observable.

If both of these conditions are satisfied, any long-distance non-perturbative corrections will be suppressed by the ratio of the long-distance scale to the short-distance one to some (observable-dependent) power, typically

$$\text{IR-safe observables: IR corrections} \;\propto\; \frac{Q_{\mathrm{IR}}^2}{Q_{\mathrm{UV}}^2} \tag{58}$$

where Q_{UV} denotes a generic hard scale in the problem, and $Q_{\mathrm{IR}} \sim \Lambda_{\mathrm{QCD}} \sim \mathcal{O}(1\ \mathrm{GeV})$.

Due to this *power suppression*, IR-safe observables are not so sensitive to our lack of ability to solve the strongly coupled IR physics, unless of course we go to processes for which the relevant hard scale, Q_{UV}, is small (such as minimum-bias, soft jets, or small-scale jet substructure). Even when a high scale is present, however, as in resonance decays, jet fragmentation or underlying-event-type studies, IR safety only guarantees us that IR corrections are small, not that they are zero. Thus, ultimately, we run into a precision barrier even for IR-safe observables, which only a reliable understanding of the long-distance physics itself can address.

To constrain models of long-distance physics, one needs IR *sensitive* observables. Hence it is not always the case that IR-safe observables are preferable—the purpose decides the tool. Instead of the suppressed corrections above, the perturbative prediction for such observables contains logarithms of the type already encountered in Eq. (46),

$$\text{IR sensitive observables: IR corrections} \;\propto\; \alpha_{\mathrm{s}}^n \log^{m\leq 2n}\left(\frac{Q_{\mathrm{UV}}^2}{Q_{\mathrm{IR}}^2}\right), \tag{59}$$

which grow increasingly large as $Q_{\mathrm{IR}}/Q_{\mathrm{UV}} \to 0$. As an example, consider such a fundamental quantity as particle multiplicities (= number of particles); in the absence of non-trivial IR effects, the number of partons tends logarithmically to infinity as the IR cutoff is lowered. Similarly, the distinction between a charged and a neutral pion only occurs in the very last phase of hadronization, and hence observables that only include charged tracks, for instance, are always IR sensitive[19].

Two important categories of IR-safe observables that are widely used are *event shapes* and *jet algorithms*. Jet algorithms are perhaps nowhere as pedagogically described as in the 2009 ESHEP lectures by Salam [4, Chapter 5]. Event shapes in the context of hadron colliders have not yet been as widely explored, but the basic phenomenology is introduced also by Salam and collaborators in Ref. [82], with first measurements reported by CMS and ATLAS [83, 84] and a proposal to use them also for the characterization of soft-QCD ('minimum-bias') events put forth in Ref. [85].

Let us here merely emphasize that the real reason to prefer IR-safe jet algorithms over unsafe ones is not that they necessarily give very different or 'better' answers in the experiment—experiments are IR-safe by definition, and the difference between IR-safe and unsafe algorithms may not even be visible when running the algorithm on experimental data—but that it is only possible to compute pQCD predictions for the IR-safe ones. Any measurement performed with an IR-unsafe algorithm can only be compared to calculations that include a detailed hadronization model. This both limits the number of calculations that can be compared to and also adds an a priori unknown sensitivity to the details of the hadronization description, details which one would rather investigate and constrain separately, in the framework of more dedicated fragmentation studies.

For LHC phenomenology, the preferred IR-safe algorithm for jet reconstruction is currently the *anti-kT* one [86], with size parameter R varying between 0.4 and 0.7, although larger sizes can be motivated in certain contexts, for example, to look for highly energetic jets and/or the boosted decay products of high-mass objects [23, 87]. This algorithm generates circular-looking jets, so subtracting the energy believed to be associated with the *underlying event* (UE, see Section 5.2) is particularly simple.

[19] This remains true in principle even if the tracks are clustered into jets, although the energy clustered in this way does provide a lower bound on Q_{UV} in the given event, since 'charged + neutral > charged-only'.

For jet substructure, typically either the 'kT' or 'Cambridge/Aachen' algorithms are used, see, for example, Ref. [23, 87]. The clustering measures used in these algorithms more closely mimic the singularity structure of QCD bremsstrahlung and they are therefore particularly well suited to 'unravel' a tree of QCD branchings [4], such as those a parton shower generates. The Cambridge/Aachen algorithm may also be used to characterize the underlying event, see Ref. [88].

3 Monte Carlo event generators

In this section, we discuss the physics of Monte Carlo event generators and their mathematical foundations, at an introductory level. We shall attempt to convey the main ideas as clearly as possible without burying them in an avalanche of technical details. References to more detailed discussions are included where applicable. We assume the reader is already familiar with the contents of the preceding section on hard processes.

The task of a Monte Carlo event generator is to calculate everything that happens in a high-energy collision, from the hard short-distance physics to the long wavelengths of hadronization and hadron decays. Obviously, this requires some compromises to be made. General-purpose generators like HERWIG [39, 89], PYTHIA [90, 91], and SHERPA [92], start from low-order (LO or NLO) descriptions of the perturbative hard physics and then attempt to include the 'most significant' corrections, such as higher-order ME corrections and parton showers, resonance decays and finite-width effects, underlying event, beam remnants, hadronization and hadron decays. Each of them had slightly different origins, which carries through to the emphasis placed on various physics aspects today.

- PYTHIA. Successor to JETSET (begun in 1978). Originated in hadronization studies. Main feature: the Lund string fragmentation model.
- HERWIG. Successor to EARWIG (begun in 1984). Originated in perturbative coherence studies. Main feature: angular-ordered parton showers.
- SHERPA. Begun in 2000. Originated in studies of the matching of hard-emission MEs with parton showers. Main feature: CKKW matching.

There is also a large number of more specialized generators, mainly for hard processes within and beyond the Standard Model (SM), a few that offer alternative shower models, and ones specializing in soft-inclusive and/or heavy-ion physics.

An important aspect of contemporary generators is the ability to combine specialized ones with general-purpose ones, via interfaces. The most common interface between partonic hard-process and parton-shower generators is the Les Houches event file standard, defined in Ref. [93, 94] and 'spoken' by most modern generator tools. For interfaces to experimental analysis packages (like RIVET [95]) and detector simulations (like GEANT [96]), typically the HepMC standard is used [97].

Hard processes were the topic of Section 2. In this section, we shall focus mainly on parton showers, with some brief comments on resonance decays at the end. Section 4 then concerns the matching of MEs and parton showers. Finally, models of hadronization and the UE are the topic of Section 5.

Several of the discussions below rely on material from the section on Monte Carlo event generators in the PDG review of particle physics [24] and on the more comprehensive review by the *MCnet* collaboration in Ref. [5]. The latter also contains brief descriptions of the physics implementations of each of the main general-purpose event generators on the market, together with a guide on how to use (and not use) them in various connections, and a collection of comparisons to important experimental distributions. We highly recommend readers to obtain a copy of that review, as it is the most comprehensive and up-to-date review of event generators currently available. Another useful and pedagogical review on event generators is contained in the 2006 ESHEP lectures by Torbjörn Sjöstrand [41], with a more recent update in Ref. [98].

Table 2: Relative uncertainty after n evaluations, in one and d dimensions, for two traditional numerical integration methods and stochastic Monte Carlo. The last column shows the number of function evaluations that are required per point, in d dimensions.

Relative uncertainty with n points	**1-Dim**	**d-Dim**	**n_{eval}/point**
Trapezoidal rule	$1/n^2$	$1/n^{2/d}$	2^d
Simpson's rule	$1/n^4$	$1/n^{4/d}$	3^d
Monte Carlo	$1/\sqrt{n}$	$1/\sqrt{n}$	1

3.1 The Monte Carlo method

A ubiquitous problem in fundamental physics is the following: given a source located some distance from a detector, predict the number of counts that should be observed within the solid angle spanned by the detector (or within a bin of its phase-space acceptance), as a function of the properties of the source, the intervening medium and the efficiency of the detector. Essentially, the task is to compute integrals of the form

$$N_{\text{Count}}(\Delta\Omega) \;=\; \int_{\Delta\Omega} \mathrm{d}\Omega \, \frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}, \tag{60}$$

with $\mathrm{d}\sigma$ a differential cross section for the process of interest.

In particle physics, phase space has three dimensions per final-state particle (minus four for overall four-momentum-conservation). Thus, for problems with more than a few outgoing particles, the dimensionality of phase space increases rapidly. At LEP, for instance, the total multiplicity of neutral + charged hadrons (before weak decays) was typically ~ 30 particles, for about 86 dimensions.

The standard one-dimensional numerical-integration methods give very slow convergence rates for higher-dimensional problems. For illustration, a table of convergence rates in one and d dimensions is given in Table 2, comparing the trapezoidal (2-point) rule and Simpson's (3-point) rule to random-number-based Monte Carlo. In one dimension, the $1/n^2$ convergence rate of the trapezoidal rule is much faster than the stochastic $1/\sqrt{n}$ of random-number Monte Carlo, and Simpson's rule converges even faster. However, as we go to d dimensions, the convergence rate of the n-point rules all degrade with d (while the number of function evaluations required for each 'point' simultaneously increases). The MC convergence rate, on the other hand, remains the simple stochastic $1/\sqrt{n}$, independent of d, and each point still only requires one function evaluation. These are some of the main reasons that MC is the preferred numerical integration technique for high-dimensional problems. In addition, the random phase-space vectors it generates can be re-used in many ways, for instance as an input to iterative solutions, to compute many different observables simultaneously and/or to hand 'events' to propagation and detector-simulation codes.

Therefore, virtually all numerical cross-section calculations are based on Monte Carlo techniques in one form or another, the simplest being the RAMBO algorithm [99] which can be expressed in about half a page of code and generates a flat scan over n-body phase space[20].

However, due to the IR singularities in pQCD, and due to the presence of short-lived resonances, the functions to be integrated, $|\mathcal{M}_{F+k}|^2$, can be highly non-uniform, especially for large k. This implies that we will have to be clever in the way we sample phase space if we want the integration to converge in any reasonable amount of time—simple algorithms like RAMBO quickly become inefficient for k greater than a few. To address this bottleneck, the simplest step up from RAMBO is to introduce generic (i.e., automated) importance-sampling methods, such as those offered by the VEGAS algorithm [100, 101]. This is still the dominant basic technique, although most modern codes do employ several additional refinements, such as several different copies of VEGAS running in parallel (multi-channel integration),

[20] Strictly speaking, RAMBO is only truly uniform for massless particles. Its massive variant makes up for phase-space biases by returning weighted momentum configurations.

"This risk, that convergence is only given with a certain probability, is inherent in Monte Carlo calculations and is the reason why this technique was named after the world's most famous gambling casino. Indeed, the name is doubly appropriate because the style of gambling in the Monte Carlo casino, not to be confused with the noisy and tasteless gambling houses of Las Vegas and Reno, is serious and sophisticated."

F. James, "Monte Carlo theory and practice", Rept. Prog. Phys. *43 (1980) 1145*

Fig. 15: *Left:* The casino in Monaco. *Right:* extract from Ref. [7] concerning the nature of Monte Carlo techniques.

to further optimize the sampling. Alternatively, a few algorithms incorporate the singularity structure of QCD explicitly in their phase-space sampling, either by directly generating momenta distributed according to the leading-order QCD singularities, in a sort of 'QCD-preweighted' analogue of RAMBO, called SARGE [102], or by using all-orders Markovian parton showers to generate them (VINCIA [80, 81]).

The price of using random numbers is that we must generalize our notion of convergence. In calculus, we say that a sequence $\{A\}$ *converges* to B if an n exists for which the difference $|A_{i>n} - B| < \epsilon$ for any $\epsilon > 0$. In random-number-based techniques, we cannot completely rule out the possibility of very pathological sequences of 'dice rolls' leading to large deviations from the true goal, and hence we are restricted to say that $\{A\}$ converges to B if an n exists for which *the probability* for $|A_{i>n} - B| < \epsilon$, for any $\epsilon > 0$, is greater than P, for any $P \in [0, 1]$ [7]. This risk, that convergence is only given with a certain probability, is the reason why Monte Carlo techniques were named after the famous casino in Monaco, illustrated in Fig. 15.

3.2 Theoretical basis of parton showers

In Section 2, we noted two conditions that had to be valid for fixed-order truncations of the perturbative series to be valid. Firstly, the strong coupling α_s must be small for perturbation theory to be valid at all. This restricts us to the region in which all scales $Q_i \gg \Lambda_{\rm QCD}$. We shall maintain this restriction in this section, that, we are still considering *pQCD*. Secondly, however, in order to be allowed to *truncate* the perturbative series, we had to require $\sigma_{k+1} \ll \sigma_k$, that is, the corrections at successive orders must become successively smaller, which—due to the enhancements from soft/collinear singular (conformal) dynamics—effectively restricted us to consider only the phase-space region in which all jets are 'hard and well-separated', equivalent to requiring all $Q_i/Q_j \approx 1$. In this section, we shall see how to lift this restriction, extending the applicability of perturbation theory into regions that include scale hierarchies, $Q_i \gg Q_j \gg \Lambda_{\rm QCD}$, such as those which occur for soft jets, jet substructure, etc.

In fact, the simultaneous restriction to all resolved scales being larger than $\Lambda_{\rm QCD}$ *and* no large hierarchies is extremely severe, if taken at face value. Since we collide and observe hadrons ($\rightarrow$ low scales) while simultaneously wishing to study short-distance physics processes ($\rightarrow$ high scales), it would appear trivial to conclude that fixed-order pQCD is not applicable to collider physics at all. So why do we still use it?

The answer lies in the fact that we actually never truly perform a fixed-order calculation in QCD. Let us repeat the factorized formula for the cross section, Eq. (44), now inserting also a function, D, to

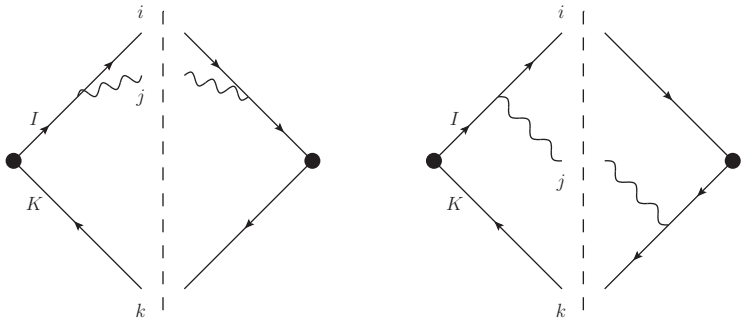


Fig. 16: Diagrams (squared) giving rise to collinear (*left*) and soft (*right*) singularities

represent the fragmentation of the final-state partons into observable hadrons,

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\mathcal{O}} = \sum_{i,j} \int_0^1 \mathrm{d}x_i\, \mathrm{d}x_j \sum_f \int \mathrm{d}\Phi_f \; f_{i/h_1}(x_i,\mu_F^2) f_{j/h_2}(x_j,\mu_F^2) \frac{\mathrm{d}\hat{\sigma}_{ij\to f}}{\mathrm{d}\hat{\mathcal{O}}} D_f(\hat{\mathcal{O}} \to \mathcal{O}, \mu_F^2)\,, \qquad (61)$$

with $\hat{\mathcal{O}}$ denoting the observable evaluated on the partonic final state, and $\mathcal{O}$ the observable evaluated on the hadronic final state, after fragmentation. Although the partonic cross section, $\mathrm{d}\hat{\sigma}_{ij\to f}$, does represent a fixed-order calculation, the parton densities, f_{i/h_1} and f_{j/h_2}, include so-called resummations of perturbative corrections *to all orders* from the initial scale of the order of the mass of the proton, up to the factorization scale, μ_F (see Section 2.2 and/or the TASI lectures by Sterman [43]). Note that the oft-stated mantra that the PDFs are purely non-perturbative functions is therefore misleading. True, they are defined as essentially non-perturbative functions at some very low scale, $\mu_0 \sim$ a few GeV, but, if μ_F is taken large, they necessarily incorporate a significant amount of perturbative physics as well. On the 'fixed-order side', all we have left to ensure in $\mathrm{d}\sigma_{ij\to f}$ is then that there are no large hierarchies remaining between μ_F and the QCD scales appearing in Φ_f. Likewise, in the final state, the fragmentation functions, D_f, include infinite-order resummations of perturbative corrections all the way *from* μ_F down to some low scale, with similar caveats concerning mantras about their non-perturbative nature as for the PDFs.

3.2.1 Step one: Infinite legs

The infinite-order resummations that are included in objects such as the PDFs and fragmentation functions in Eq. (61) (and in their parton-shower equivalents) rely on some very simple and powerful properties of gauge-field theories that were already touched on in Section 2. In particular, we saw in Section 2.4 that we can represent all the IR limits of any NLO amplitude with a set of simple universal functions, based solely on knowing which partons are colour-connected (i.e., have colour-space index contractions) with one another.

The diagrams in Fig. 16 show the basic origin of the universal IR singularities of gauge-theory amplitudes. On the left is shown a diagram (squared) in which an emission with small s_{ij} interferes with itself. In the collinear limit, $s_{ij} \to 0$, the propagator of the parent parton, I, goes on shell; the singularity of the associated propagator factor is the origin of the $1/s_{ij}$ collinear singularities. On the right is shown the interference between a diagram with emission from parton I and a diagram with emission from parton K. The resulting term has propagator singularities when both partons I and K go on shell, which can happen simultaneously if parton j is soft. This generates the $2s_{ik}/(s_{ij}s_{jk})$ soft singularity, also called the soft-eikonal factor or the dipole factor.

We now understand the fundamental origin of the IR singularities, why they are universal, and why amplitudes factorize in the soft and collinear limits—the singularities are simply generated by inter-

Fig. 17: Illustration of the branching phase space for $\mathrm{q\bar{q}} \to \mathrm{qg\bar{q}}$, with the original dipole-antenna oriented horizontally, the two parents sharing the transverse component of recoil, and the azimuthal angle ϕ (representing rotations of the emitted parton around the dipole axis) chosen such that the gluon is radiated upwards. From Ref. [80].

mediate parton propagators going on shell, which is independent of the nature of the hard process, and hence can be factorized from it.

Thus, for each pair of (massless) colour-connected partons I and K in F, the squared amplitude for $F+1$ gluon, $|\mathcal{M}_{F+1}|^2$, will include a factor

$$|\mathcal{M}_{F+1}|^2 \;=\; \underbrace{g_{\mathrm{s}}^2\, N_{\mathrm{C}} \left(\frac{2 s_{ik}}{s_{ij} s_{jk}} \;+\; \text{collinear terms} \right)}_{\text{antenna function}} |\mathcal{M}_F|^2 \,, \tag{62}$$

where $g_{\mathrm{s}}^2 = 4\pi\alpha_{\mathrm{s}}$ is the strong coupling, i and k represent partons I and K after the branching (i.e., they include possible recoil effects) and s_{ij} is the invariant between parton i and the emitted parton, j.

The branching phase space of a colour dipole (i.e., a pair of partons connected by a colour-index contraction) is illustrated in Fig. 17. Expressed in the branching invariants, s_{ij} and s_{jk}, the phase space has a characteristic triangular shape, imposed by the relation $s = s_{ij} + s_{jk} + s_{ik}$ (assuming massless partons). Sketchings of the post-branching parton momenta have been inserted in various places in the figure, for illustration. The soft singularity is located at the origin of the plot and the collinear regions lie along the axes.

The collinear terms for a $\mathrm{q\bar{q}} \to \mathrm{qg\bar{q}}$ 'antenna' are unambiguous and are given in Section 2.4. Since gluons are in the adjoint representation, they carry both a colour and an anticolour index (one corresponding to the rows and the other to the columns of the Gell–Mann matrices), and there is therefore some ambiguity concerning how to partition collinear radiation among the two antennae they participate in. This is discussed in more detail in Ref. [80]. Differences are subleading, however, and for our purposes here we shall consider gluon antenna ends as radiating just like quark ones. The difference between quark and gluon radiation then arise mainly because gluons participate in two antennae, while quarks only participate in one. This is related to the difference between the colour factors, $C_A \sim 2C_F$.

The problem that plagued the fixed-order truncations in Section 2 is clearly visible in Eq. (62): if we integrate over the entire phase space including the region $s_{ij} \to 0$, $s_{jk} \to 0$, we end up with a double pole. If we instead regulate the divergence by cutting off the integration at some minimal *perturbative cutoff scale* μ_{IR}^2, we end up with a logarithm squared of that scale. This is a typical example of 'large logarithms' being generated by the presence of scale hierarchies. Also note that the precise definition of μ_{IR} is not unique. Any scale choice that properly isolates the singularities from the rest of phase space will do, with some typical choices being, for example, invariant-mass and/or transverse-momentum scales.

Before we continue, it is worth noting that Eq. (62) is often rewritten in other forms to emphasize specific aspects of it. One such rewriting is thus to reformulate the invariants s_{ij} appearing in it in terms of energies and angles,

$$s_{ij} = 2E_i E_j \left(1 - \cos\theta_{ij}\right). \tag{63}$$

Rewritten in this way, the differentials can become partial fractions,

$$\frac{\mathrm{d}s_{ij}}{s_{ij}} \frac{\mathrm{d}s_{jk}}{s_{jk}} \propto \frac{\mathrm{d}E_j}{E_j} \frac{\mathrm{d}\theta_{ij}}{\theta_{ij}} + \frac{\mathrm{d}E_j}{E_j} \frac{\mathrm{d}\theta_{jk}}{\theta_{jk}}. \tag{64}$$

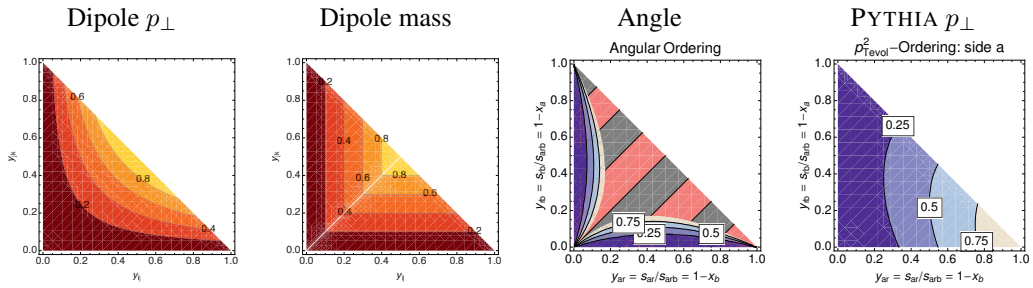


Fig. 18: A selection of parton-shower evolution variables, represented as contours over the dipole phase space. Note: the right-most variable corresponds to evolution of only one of the parents, the one with no collinear singularity along the bottom of the plot.

This kind of rewriting enables an intuitively appealing categorization of the singularities as related to vanishing energies and angles, explaining why they are called *soft* and *collinear*, respectively. Arguments based on this rewriting have led to important insights in QCD. For instance, within the framework of conventional parton showers, it was shown in a sequence of publications (see Refs. [103, 104] and references therein) that the destructive interference effects between two or more colour-connected partons (*coherence*) can be described by using the angle of the emissions as the shower-ordering variable. One should still keep in mind, however, that Lorentz-non-invariant formulations come with similar caveats and warnings as do gauge-non-invariant formulations of quantum field theory: while they can be practical to work with at intermediate stages of a calculation, one should be careful with any physical conclusions that rely explicitly on them.

We shall therefore here restrict ourselves to a Lorentz-invariant formalism based directly on Eq. (62), pioneered by the dipole formulation of QCD cascades [76]. The collinear limit is then replaced by a more general *single-pole* limit in which a single parton–parton invariant vanishes (as, *for instance*, when a pair of partons become collinear), while the soft limit is replaced by one in which two (or more) invariants involving the same parton vanish simultaneously (as, for instance by that parton becoming soft in a frame defined by two or more hard partons). This avoids frame-dependent ambiguities from entering into the language, at the price of a slight reinterpretation of what is meant by collinear and soft.

In the generator landscape, *angular ordering* is used by the HERWIG [104] and HERWIG++ [105] programs, and an *angular veto* is imposed on the virtuality-ordered evolution in PYTHIA 6 [106]. Variants of the dipole approach is used by the ARIADNE [107], SHERPA [108, 109], and VINCIA [110] programs, while the $p_\perp$-ordered showers in PYTHIA 6 and 8 represent a hybrid, combining collinear splitting kernels with dipole kinematics [111]. Phase-space contours of equal value of some of these choices are illustrated in Fig. 18. During the shower evolution, each model effectively 'sweeps' over phase space in the order implied by these contours. For example, a $p_\perp$-ordered dipole shower (left-most plot in Fig. 18) will treat a hard-collinear branching as occurring 'earlier' than a soft one, while a mass-ordered dipole shower (second plot) will tend to do the opposite. This affects the tower of virtual corrections generated by each shower model via the so-called Sudakov factor, discussed below. Experimental tests of the subleading aspects of shower models can therefore be quite important, see Ref. [112] for a recent example.

Independently of rewritings and philosophy, the real power of Eq. (62) lies in the fact that it is *universal*. Thus, for *any* process F, we can apply Eq. (62) in order to get an approximation for $\mathrm{d}\sigma_{F+1}$. We may then, for instance, take our newly obtained expression for $F+1$ as our arbitrary process and crank Eq. (62) again, to obtain an approximation for $\mathrm{d}\sigma_{F+2}$, and so forth. What we have here is therefore a very simple recursion relation that can be used to generate approximations to leading-order cross sections with arbitrary numbers of additional legs. The quality of this approximation is governed by how many terms besides the leading one shown in Eq. (53) are included in the game. Including all possible terms,

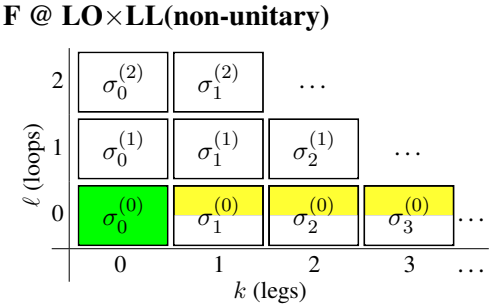


Fig. 19: Coefficients of the perturbative series covered by LO + LL approximations to higher-multiplicity tree-level ME. Green (darker) shading represents the full perturbative coefficient at the respective k and ℓ. Yellow (lighter) shading represents an LL approximation to it. Half-shaded boxes indicate phase spaces in which we are prohibited from integrating over the IR singular region, as discussed in Sections 2.3 and 4.

the most general form for the cross section at $F+n$ jets, restricted to the phase-space region above some IR-cutoff scale $\mu_{\rm IR}$, has the following algebraic structure,

$$\sigma^{(0)}_{F+n} = \alpha_{\rm s}^n \left(\ln^{2n} + \ln^{2n-1} + \ln^{2n-2} + \cdots + \ln\, + \mathcal{F}\right) \tag{65}$$

where we use the notation $\ln^\lambda$ without an argument to denote generic functions of *transcendentality* λ (the logarithmic function to the power λ being a 'typical' example of a function with transcendentality λ appearing in cross section expressions, but also dilogarithms and higher logarithmic functions[21] of transcendentality > 1 should be implicitly understood to belong to our notation $\ln^\lambda$). The last term, $\mathcal{F}$, represents a rational function of transcendentality, 0. We shall also use the nomenclature *singular* and *finite* for the $\ln^\lambda$ and $\mathcal{F}$ terms, respectively, a terminology which reflects their respective behaviour in the limit $\mu_{\rm IR} \to 0$.

The simplest approximation one can build on Eq. (65), dropping all but the leading $\ln^{2n}$ term in the parenthesis, is thus the *leading-transcendentality* approximation. This approximation is better known as the double logarithmic approximation, since it generates the correct coefficient for terms which have two powers of logarithms for each power of $\alpha_{\rm s}$, while terms of lower transcendentalities are not guaranteed to have the correct coefficients. In so-called LL parton-shower algorithms, one generally expects to reproduce the correct coefficients for the $\ln^{2n}$ and $\ln^{2n-1}$ terms. In addition, several formally subleading improvements are normally also introduced in such algorithms (such as explicit momentum conservation, gluon polarization and other spin-correlation effects [113–115], higher-order coherence effects [103], renormalization scale choices [116], finite-width effects [117], etc), as a means to improve the agreement with some of the more subleading coefficients as well, if not in every phase-space point then at least on average. Although LL showers do not magically acquire NLL (next-to-leading-log) precision from such procedures, one, therefore, still expects a significantly better average performance from them than from corresponding 'strict' LL analytical resummations. A side effect of this is that it is often possible to 'tune' shower algorithms to give better-than-nominal agreement with experimental distributions, by adjusting the parameters controlling the treatment of subleading effects. One should remember, however, that there is a limit to how much can be accomplished in this way—at some point, agreement with one process will only come at the price of disagreement with another, and at this point further tuning would be meaningless.

Applying such an iterative process on a Born-level cross section, one obtains the description of the full perturbative series illustrated in Fig. 19. The yellow (lighter) shades, used here for $k \geq 1$, indicate that the coefficient obtained is not the exact one, but rather an approximation to it that only

[21] Note: due to the theorems that allow us, for instance, to rewrite dilogarithms in different ways with logarithmic and lower 'spillover' terms, the coefficients at each λ are only well-defined up to reparameterization ambiguities involving the terms with transcendentality greater than λ.

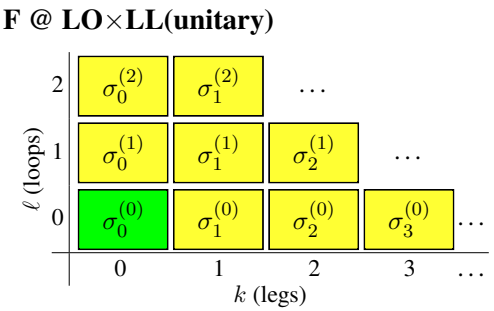


Fig. 20: Coefficients of the perturbative series covered by LO + LL calculations, imposing unitarity order by order for each $n = k + \ell$. Green (darker) shading represents the full perturbative coefficient at the respective k and ℓ. Yellow (lighter) shading represents an LL approximation to it.

gets its leading singularities right. However, since this is still only an approximation to infinite-order *tree-level* cross sections (we have not yet included any virtual corrections), we cannot yet integrate this approximation over all of phase space, as illustrated by the yellow boxes being only half filled in Fig. 19; otherwise, the summed-total cross section would still be infinite. This particular approximation would therefore still appear to be very useless indeed—on one hand, it is only guaranteed to get the singular terms right, but on the other, it does not actually allow us to integrate over the singular region. In order to obtain a truly *all-orders* calculation, the constraint of unitarity must also be explicitly imposed, which furnishes an approximation to all-orders loop corrections as well. Let us therefore emphasize that Fig. 19 is included for pedagogical purposes only; all resummation calculations, whether analytical or parton-shower based, include virtual corrections as well and consequently yield finite total cross sections, as will now be described.

3.2.2 Step two: Infinite loops

Order-by-order unitarity, such as used in the KLN theorem, implies that the singularities caused by integration over unresolved radiation in the tree-level MEs must be cancelled, order by order, by equal but opposite-sign singularities in the virtual corrections at the same order. That is, from Eq. (62), we immediately know that the one-loop correction to $\mathrm{d}\sigma_F$ *must* contain a term,

$$2\mathrm{Re}[\mathcal{M}_F^{(0)}\mathcal{M}_F^{(1)*}] \supset -g_\mathrm{s}^2\, N_\mathrm{C} \left|\mathcal{M}_F^{(0)}\right|^2 \int \frac{\mathrm{d}s_{ij}\,\mathrm{d}s_{jk}}{16\pi^2 s_{ijk}} \left(\frac{2s_{ik}}{s_{ij}s_{jk}} + \text{less singular terms}\right), \qquad (66)$$

that cancels the divergence coming from Eq. (62) itself. Further, since this is universally true, we may apply Eq. (66) again to get an approximation to the corrections generated by Eq. (62) at the next order and so on. By adding such terms explicitly, order by order, we may now bootstrap our way around the entire perturbative series, using Eq. (62) to move horizontally and Eq. (66) to move along diagonals of constant $n = k + \ell$. Since real-virtual cancellations are now explicitly restored, we may finally extend the integrations over all of phase space, resulting in the picture shown in Fig. 20.

The picture shown in Fig. 20, not the one in Fig. 19, corresponds to what is actually done in *resummation* calculations, both of the analytic and parton-shower types[22]. Physically, there is a significant and intuitive meaning to the imposition of unitarity, as follows.

Take a jet algorithm, with some measure of jet resolution, Q, and apply it to an arbitrary sample of events, say dijets. At a very crude resolution scale, corresponding to a high value for Q, you find that

[22] In the way these calculations are formulated in practice, they in fact rely on one additional property, called exponentiation, that allows us to move along straight vertical lines in the loops-and-legs diagrams. However, since the two different directions furnished by Eqs. (62) and (66) are already sufficient to move freely in the full two-dimensional coefficient space, we shall use exponentiation without extensively justifying it here.

everything is clustered back to a dijet configuration, and the two-jet cross section is equal to the total inclusive cross section,

$$\sigma_{\rm tot} = \sigma_{F;\rm incl} \ . \tag{67}$$

At finer resolutions, decreasing Q, you see that some events that were previously classified as two-jet events contain additional, lower-scale jets, that you can now resolve, and hence those events now migrate to the three-jet bin, while the total inclusive cross section of course remains unchanged,

$$\sigma_{\rm tot} = \sigma_{F;\rm incl} = \sigma_{F;\rm excl}(Q) + \sigma_{F+1;\rm incl}(Q), \tag{68}$$

where 'incl' and 'excl' stands for inclusive and exclusive cross sections[23], respectively, and the Q-dependence in the two terms on the right-hand side must cancel so that the total inclusive cross section is independent of Q. Later, some three-jet events now migrate further, to four and higher jets, while still more two-jet events migrate *into* the three-jet bin, etc. For arbitrary n and Q, we have

$$\sigma_{F+n;\rm incl}(Q) = \sigma_{F;\rm incl} - \sum_{m=0}^{n-1} \sigma_{F+m;\rm excl}(Q). \tag{69}$$

This equation expresses the trivial fact that the cross section for n or more jets can be computed as the total inclusive cross section for F minus a sum over the cross sections for F + exactly m jets including all $m < n$. On the theoretical side, it is these negative terms which must be included in the calculation, for each order $n = k + \ell$, to restore unitarity. Physically, they express that, at a given scale Q, each event will be classified as having *either* zero, one, two, or whatever jets. Or, equivalently, for each event we gain in the three-jet bin as Q is lowered, we must loose one event in the two-jet one; the negative contribution to the two-jet bin is exactly minus the integral of the positive contribution to the three-jet one, and so on. We may perceive this *detailed balance* as an *evolution* of the event structure with Q, for each event, which is effectively what is done in parton-shower algorithms, to which we shall return in Section 3.3.

3.3 Perturbation theory with Markov chains

Consider again the Born-level cross section for an arbitrary hard process, F, differentially in an arbitrary IR-safe observable $\mathcal{O}$, as obtained from Eq. (45):

$$\left.\frac{{\rm d}\sigma_F^{(0)}}{{\rm d}\mathcal{O}}\right|_{\rm Born} = \int {\rm d}\Phi_F \ |\mathcal{M}_F^{(0)}|^2 \, \delta(\mathcal{O} - \mathcal{O}(\Phi_F)), \tag{70}$$

where the integration runs over the full final-state on-shell phase space of F (this expression and those below would also apply to hadron collisions were we to include integrations over the PDFs in the initial state), and the δ function projects out a one-dimensional slice defined by $\mathcal{O}$ evaluated on the set of final-state momenta which we denote Φ_F.

To make the connection to parton showers, we insert an operator, $\mathcal{S}$, that acts on the Born-level final state *before* the observable is evaluated, that is,

$$\left.\frac{{\rm d}\sigma_F}{{\rm d}\mathcal{O}}\right|_{\mathcal{S}} = \int {\rm d}\Phi_F \ |\mathcal{M}_F^{(0)}|^2 \, \mathcal{S}(\Phi_F, \mathcal{O}). \tag{71}$$

Formally, this operator—the evolution operator—will be responsible for generating all (real and virtual) higher-order corrections to the Born-level expression. The measurement δ function appearing explicitly in Eq. (70) is now implicit in $\mathcal{S}$.

[23] F *inclusive* $= F$ plus anything. F *exclusive* $= F$ and only F. Thus, $\sigma_{F;\rm incl} = \sum_{\rm k=0}^{\infty} \sigma_{F+k;\rm excl}$

Algorithmically, parton showers cast $\mathcal{S}$ as an iterative Markov (i.e., history-independent) chain, with an evolution parameter, $Q_{\rm E}$, that formally represents the factorization scale of the event, below which all structure is summed over inclusively. Depending on the particular choice of shower algorithm, $Q_{\rm E}$ may be defined as a parton virtuality (virtuality-order showers), as a transverse-momentum scale ($p_\perp$-ordered showers) or as a combination of energies times angles (angular ordering). Regardless of the specific form of $Q_{\rm E}$, the evolution parameter will go towards zero as the Markov chain develops, and the event structure will become more and more exclusively resolved. A transition from a perturbative evolution to a non-perturbative one can also be inserted, when the evolution reaches an appropriate scale, typically around 1 GeV. This scale, called the *hadronization scale*, thus represents the lowest perturbative scale that can appear in the calculations, with all perturbative corrections below it summed over inclusively.

Working out the precise form that $\mathcal{S}$ must have in order to give the correct expansions discussed in Section 3.2 takes a bit of algebra and is beyond the scope we aim to cover in these lectures. Heuristically, the procedure is as follows. We noted that the singularity structure of QCD is universal and that at least its first few terms are known to us. We also saw that we could iterate that singularity structure, using universality and unitarity, thereby bootstrapping our way around the entire perturbative series. This was illustrated by Fig. 20 in Section 3.2.

Skipping intermediate steps, the form of the all-orders pure-shower Markov chain, for the evolution of an event between two scales $Q_1 > Q_{\rm E} > Q_2$, is,

$$\begin{aligned}\mathcal{S}(\Phi_F, Q_1, Q_2, \mathcal{O}) = &\underbrace{\Delta(\Phi_F, Q_1, Q_2)\,\delta\left(\mathcal{O} - \mathcal{O}(\Phi_F)\right)}_{F+0 \text{ exclusive above } Q_2} \\ &\underbrace{+\sum_r \int_{Q_{E2}}^{Q_{E1}} \frac{\mathrm{d}\Phi^r_{F+1}}{\mathrm{d}\Phi_F}\, S_{\rm r}(\Phi_{F+1})\,\Delta(\Phi_F, Q_1, Q_{F+1})\,\mathcal{S}(\Phi_{F+1}, Q_{F+1}, Q_2, \mathcal{O})}_{F+1 \text{ inclusive above } Q_2}\,,\end{aligned} \tag{72}$$

with the so-called *Sudakov factor*,

$$\Delta(\Phi_F, Q_1, Q_2) = \exp\left[-\sum_r \int_{Q_2}^{Q_1} \frac{\mathrm{d}\Phi^r_{F+1}}{\mathrm{d}\Phi_F} S_{\rm r}(\Phi_{F+1})\right], \tag{73}$$

defining the probability that there is *no evolution* (i.e., no emissions) between the scales Q_1 and Q_2, according to the *radiation functions* $S_{\rm r}$ to which we shall return below. The term on the first line of Eq. (72) thus represents all events that *did not* evolve as the resolution scale was lowered from Q_1 to Q_2, while the second line contains a sum and phase-space integral over those events that *did* evolve—including the insertion of $\mathcal{S}(\Phi_{F+1})$ representing the possible further evolution of the event and completing the iterative definition of the Markov chain.

The factor $\mathrm{d}\Phi^r_{F+1}/\mathrm{d}\Phi_F$ defines the chosen phase space factorization. Our favourite is the so-called dipole-antenna factorization, whose principal virtue is that it is the simplest Lorentz-invariant factorization which is simultaneously exact over all of phase space while only involving on-shell momenta. For completeness, its form is

$$\frac{\mathrm{d}\Phi^r_{F+1}}{\mathrm{d}\Phi_F} = \frac{\mathrm{d}\Phi^r_3}{\mathrm{d}\Phi_2} = \mathrm{d}s_{a1}\,\mathrm{d}s_{1b}\,\frac{\mathrm{d}\phi}{2\pi}\,\frac{1}{16\pi^2 s_r}, \tag{74}$$

which involves just one colour–anticolour pair for each r, with invariant mass squared $s_r = (p_a + p_1 + p_b)^2$. Other choices, such as purely collinear ones (only exact in the collinear limit *or* involving explicitly off-shell momenta), more global ones involving all partons in the event (more complicated, in our opinion), or less global ones with a single parton playing the dominant role as emitter, are also possible, again depending on the specific algorithm considered.

The radiation functions S_{r} obviously play a crucial role in these equations, driving the emission probabilities. For example, if $S_{\mathrm{r}} \to 0$, then $\Delta \to \exp(0) = 1$ and all events stay in the top line of Eq. (72). Thus, in regions of phase space where S_{r} is small, there is little or no evolution. Conversely, for $S_{\mathrm{r}} \to \infty$, we have $\Delta \to 0$, implying that *all* events evolve. One possible choice for the radiation functions S_{r} was implicit in Eq. (62), in which we took them to include only the leading (double) singularities, with r representing colour–anticolour pairs. In general, the shower may exponentiate the entire set of universal singular terms, or only a subset of them (for example, the terms leading in the number of colours N_{C}), which is why we here let the explicit form of S_{r} be unspecified. Suffice it to say that in traditional parton showers, S_{r} would simply be the DGLAP splitting kernels (see, e.g., Ref. [3]), while they would be so-called dipole or antenna radiation functions in the various dipole-based approaches to QCD (see, e.g., Ref. [64, 70, 76, 80, 81, 109]).

The procedure for how to technically 'construct' a shower algorithm of this kind, using random numbers to generate scales distributed according to Eq. (72), is described more fully in Ref. [80], using a notation that closely parallels the one used here. The procedure is also described at a more technical level in the review [5], although using a slightly different notation. Finally, a pedagogical introduction to Monte Carlo methods in general can be found in Ref. [7].

3.4 Decays of unstable particles

In most BSM processes and some SM ones, an important aspect of the event simulation is how decays of short-lived particles, such as top quarks, Electro-Weak and Higgs bosons and new BSM resonances, are handled. We here briefly summarize the spectrum of possibilities, but emphasize that there is no universal standard. Users are advised to check whether the treatment of a given code is adequate for the physics study at hand.

The appearance of an unstable resonance as a physical particle at some intermediate stage of the event generation implies that its production and decay processes are treated as being factorized. This is called the *narrow width approximation* and is valid up to corrections of order Γ/m_0, with Γ the width and m_0 the pole mass of the particle. States whose widths are a substantial fraction of their mass should not be treated in this way, but rather as intrinsically virtual internal propagator lines.

For states treated as physical particles, two aspects are relevant: the mass distribution of the decaying particle itself and the distributions of its decay products. For the mass distribution, the simplest is to use a δ function at m_0. The next level up, typically used in general-purpose Monte Carlo, is to use a Breit–Wigner distribution (relativistic or non-relativistic), which formally resums higher-order virtual corrections to the mass distribution. Note, however, that this still only generates an improved picture for *moderate* fluctuations away from m_0. Similarly to above, particles that are significantly off-shell (in units of Γ) should not be treated as resonant, but rather as internal off-shell propagator lines. In most Monte Carlo codes, some further refinements are also included, for instance by letting Γ be a function of m ('running widths') and by limiting the magnitude of the allowed fluctuations away from m_0. See also Ref. [118] for an elaborate discussion of the Higgs boson lineshape.

For the distributions of the decay products, the simplest treatment is again to assign them their respective m_0 values, with a uniform (i.e., isotropic, or 'flat') phase-space distribution. A more sophisticated treatment distributes the decay products according to the differential decay MEs, capturing at least the internal dynamics and helicity structure of the decay process, including Einstein-Podolsky-Rosen(EPR)-like correlations. Further refinements include polarizations of the external states [113–115] (see also Refs. [119–121] for phenomenological studies) and assigning the decay products their own Breit–Wigner distributions, the latter of which opens the possibility to include also intrinsically off-shell decay channels, like $\mathrm{H} \to \mathrm{WW}^*$. Please refer to the physics manual of the code you are using and/or make simple cross checks such as plotting the distribution of the phase-space invariants it produces.

During subsequent showering of the decay products, most parton-shower models will preserve the

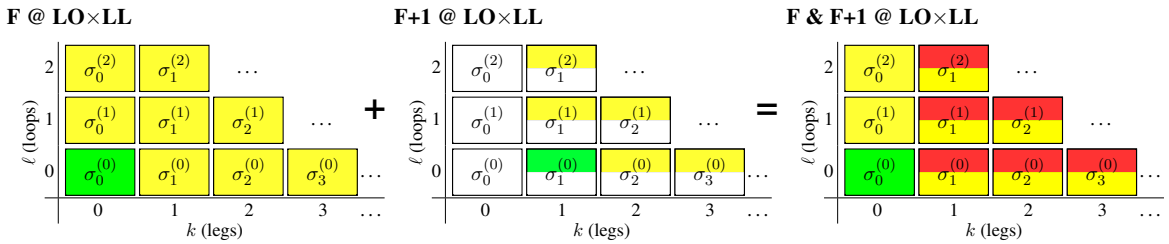


Fig. 21: The double-counting problem caused by naively adding cross sections involving MEs with different numbers of legs.

total invariant mass of each resonance-decay system separately, so as not to skew the original resonance shape.

4 Matching at LO and NLO

The essential problem that leads to ME/parton-shower matching can be illustrated in a very simple way. Assume we have computed the LO cross section for some process, F, and that we have added an LL shower to it, as in the left-hand pane of Fig. 21. We know that this only gives us an LL description of $F + 1$. We now wish to improve this from LL to LO by adding the actual LO ME for $F + 1$. Since we also want to be able to hadronize these events, etc., we again add an LL shower off them. However, since the ME for $F + 1$ is divergent, we must restrict it to cover only the phase-space region with at least one hard resolved jet, illustrated by the half-shaded boxes in the middle pane of Fig. 21.

Adding these two samples, however, we end up counting the LL terms of the inclusive cross section for $F + 1$ twice, since we are now getting them once from the shower off F and once from the ME for $F + 1$, illustrated by the dark shaded (red) areas of the right-hand pane of Fig. 21. This *double-counting* problem would grow worse if we attempted to add more MEs, with more legs. The cause is very simple. Each such calculation corresponds to an *inclusive* cross section, and hence naive addition would give

$$\sigma_{\text{tot}} = \sigma_{0;\text{incl}} + \sigma_{1;\text{incl}} = \sigma_{0;\text{excl}} + 2\,\sigma_{1;\text{incl}} \,. \tag{75}$$

Recall the definition of inclusive and exclusive cross sections, Eq. (68): F *inclusive* $= F$ plus anything. F *exclusive* $= F$ and only F. Thus, $\sigma_{F;\text{incl}} = \sum_{\text{k=0}}^{\infty} \sigma_{F+k;\text{excl}}$.

Instead, we must *match* the coefficients calculated by the two parts of the full calculation—showers and matrix elements—more systematically, for each order in perturbation theory, so that the nesting of inclusive and exclusive cross sections is respected without overcounting.

Given a parton shower and a ME generator, there are fundamentally three different ways in which we can consider matching the two [80]: slicing, subtraction and unitarity. The following subsections will briefly introduce each of these.

4.1 Slicing

The most commonly encountered matching type is currently based on separating (slicing) phase space into two regions, one of which is supposed to be mainly described by hard MEs and the other of which is supposed to be described by the shower. This type of approach was first used in HERWIG [89], to include ME corrections for one emission beyond the basic hard process [122, 123]. This is illustrated in Fig. 22. The method has since been generalized by several independent groups to include arbitrary numbers of additional legs, the most well-known of these being the CKKW [124], CKKW-L [125, 126], and MLM [127, 128] approaches.

Effectively, the shower approximation is set to zero above some scale, either due to the presence of explicit dead zones in the shower, as in HERWIG, or by vetoing any emissions above a certain *matching*

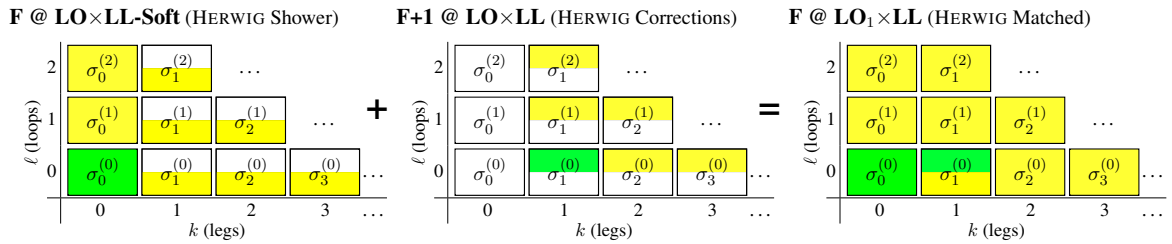


Fig. 22: HERWIG's original matching scheme [122, 123], in which the dead zone of the HERWIG shower was used as an effective 'matching scale' for one emission beyond a basic hard process.

scale, as in the CKKW(-L) and MLM approaches. The empty part of phase space can then be filled by separate events generated according to higher-multiplicity tree-level MEs. In the CKKW(-L) and MLM schemes, this process can be iterated to include arbitrary numbers of additional hard legs (the practical limit being around three or four, due to computational complexity).

In order to match smoothly with the shower calculation, the higher-multiplicity MEs must be associated with Sudakov form factors (representing the virtual corrections that would have been generated if a shower had produced the same phase-space configuration), and their α_s factors must be chosen so that, at least at the matching scale, they become identical to the choices made on the shower side [129]. The CKKW and MLM approaches do this by constructing 'fake parton-shower histories' for the higher-multiplicity MEs. By applying a sequential jet clustering algorithm, a tree-like branching structure can be created that at least has the same dominant structure as that of a parton shower. Given the fake shower tree, α_s factors can be chosen for each vertex with argument $\alpha_s(p_\perp)$ and Sudakov factors can be computed for each internal line in the tree. In the CKKW method, these Sudakov factors are estimated analytically, while the MLM and CKKW-L methods compute them numerically, from the actual shower evolution.

Thus, the matched result is identical to the ME in the region above the matching scale, modulo higher-order (Sudakov and α_s) corrections. We may sketch this as

$$\text{Matched (above matching scale)} = \overbrace{\text{Exact}}^{\text{ME}} \times \overbrace{(1+\mathcal{O}(\alpha_s))}^{\text{corrections}}, \tag{76}$$

where the 'shower corrections' include the approximate Sudakov factors and α_s reweighting factors applied to the MEs in order to obtain a smooth transition to the shower-dominated region.

Below the matching scale, the small difference between the MEs and the shower approximation can be dropped (since their leading singularities are identical and this region by definition includes no hard jets), yielding the pure shower answer in that region,

$$\begin{aligned} \text{Matched (below matching scale)} &= \overbrace{\text{Approximate}}^{\text{shower}} + \overbrace{(\text{Exact} - \text{Approximate})}^{\text{correction}} \\ &= \text{Approximate} + \text{non-singular} \\ &\to \text{Approximate}. \end{aligned} \tag{77}$$

This type of strategy is illustrated in Fig. 23.

As emphasized above, since this strategy is discontinuous across phase space, a main point here is to ensure that the behaviour across the matching scale be as smooth as possible. CKKW showed [124] that it is possible to remove any dependence on the matching scale through NLL precision by careful choices of all ingredients in the matching; technical details of the implementation (affecting the $\mathcal{O}(\alpha_s)$ terms in Eq. (76)) are important, and the dependence on the unphysical matching scale may be larger than NLL unless the implementation matches the theoretical algorithm precisely [125, 126, 130]. Furthermore,

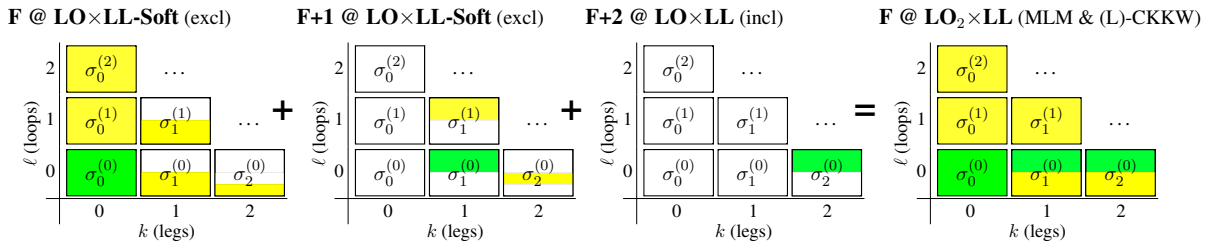


Fig. 23: Slicing, with up to two additional emissions beyond the basic process. The showers off F and $F+1$ are set to zero above a specific 'matching scale'. (The number of coefficients shown was reduced a bit in these plots to make them fit in one row.)

since the Sudakov factors are generally computed using showers (MLM, CKKW-L) or a shower-like formalism (CKKW), while the real corrections are computed using MEs, care must be taken not to (re-)introduce differences that could break the detailed real-virtual balance that ensures unitarity among the singular parts, see e.g., Ref. [129].

It is advisable not to choose the matching scale too low. This is again essentially due to the approximate scale invariance of QCD imploring us to write the matching scale as a ratio, rather than as an absolute number. If one uses a very low matching scale, the higher-multiplicity MEs will already be quite singular, leading to very large LO cross sections before matching. After matching, these large cross sections are tamed by the Sudakov factors produced by the matching scheme, and hence the final cross sections may still look reasonable. But the higher-multiplicity MEs in general contain subleading singularity structures, beyond those accounted for by the shower, and hence the delicate line of detailed balance that ensures unitarity has most assuredly been overstepped. We, therefore, recommend not to take the matching scale lower than about an order of magnitude below the characteristic scale of the hard process.

One should also be aware that all strategies of this type are quite computing intensive. This is basically due to the fact that a separate phase-space generator is required for each of the n-parton correction terms, with each such sample a priori consisting of weighted events such that a separate unweighting step (often with quite low efficiency) is needed before an unweighted sample can be produced.

4.2 Subtraction

Another way of matching two calculations is by subtracting one from the other and correcting by the difference, schematically

$$\text{Matched} = \overbrace{\text{Approximate}}^{\text{shower}} + \overbrace{(\text{Exact} - \text{Approximate})}^{\text{correction}} . \tag{78}$$

This looks very much like the structure of a subtraction-based NLO fixed-order calculation, Section 2.4, in which the shower approximation here plays the role of subtraction terms, and indeed this is what is used in strategies like MC@NLO [131–133], illustrated in Fig. 24. In this type of approach, however, negative-weight events will generally occur, for instance in phase-space points where the approximation is larger than the exact answer.

Negative weights are not in principle an insurmountable problem. Histograms can be filled with each event counted according to its weight, as usual. However, negative weights do affect efficiency. Imagine a worst-case scenario in which 1000 positive-weight events have been generated, along with 999 negative-weight ones (assuming each event weight has the same absolute value, the closest one can get to an unweighted sample in the presence of negative weights). The statistical precision of the MC answer would be equivalent to one event, for 2000 generated, i.e., a big loss in convergence rate.

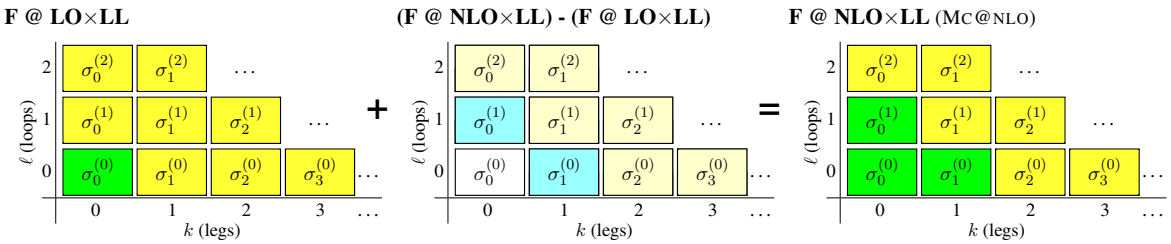


Fig. 24: MC@NLO. In the middle pane, cyan boxes denote non-singular correction terms, while the egg-coloured ones denote showers off such corrections, which cannot lead to double-counting at the LL level.

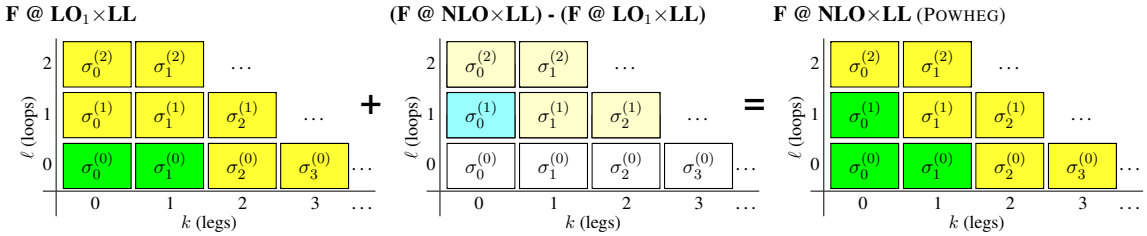


Fig. 25: POWHEG. In the middle pane, cyan boxes denote non-singular correction terms, while the egg-coloured ones denote showers off such corrections, which cannot lead to double-counting at the LL level.

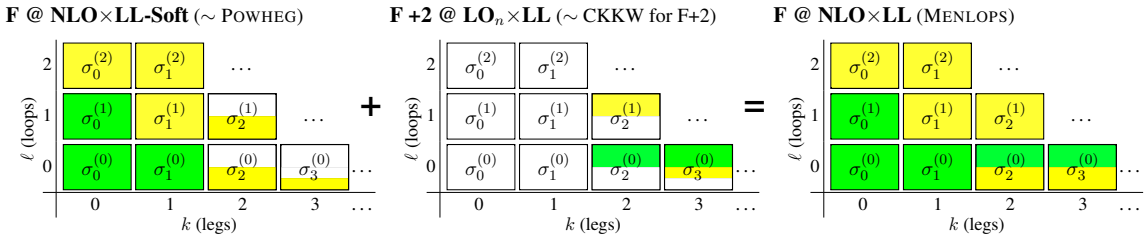


Fig. 26: MENLOPS. Note that each of the POWHEG and CKKW samples are composed of separate sub-samples, as illustrated in Figs. 23 and 25.

In practice, generators like MC@NLO 'only' produce around 10% or less events with negative weights, so the convergence rate should not be severely affected for ordinary applications. Nevertheless, the problem of negative weights motivated the development of the so-called POWHEG approach [134], illustrated in Fig. 25, which is constructed specifically to prevent negative-weight events from occurring and simultaneously to be more independent of which parton-shower algorithm it is used with. In the POWHEG method, one effectively modifies the real-emission probability for the first emission to agree with the $F + 1$ ME (this is covered under unitarity, below). One is then left with a purely virtual correction, which will typically be positive, at least for processes for which the NLO cross section is larger than the LO one.

The advantage of these methods is obviously that NLO corrections to the Born level can be systematically incorporated. However, a systematic way of extending this strategy beyond the first additional emission is not available, save for combining them with a slicing-based strategy for the additional legs, as in MENLOPS [135], illustrated in Fig. 26. These issues are, however, no more severe than in ordinary fixed-order NLO approaches, and hence they are not viewed as disadvantages if the point of reference is an NLO computation.

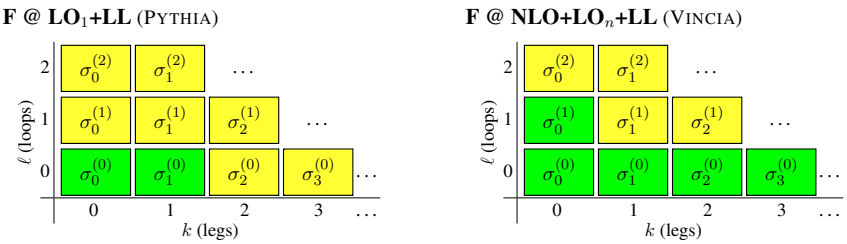


Fig. 27: PYTHIA (left) and VINCIA (right). Unitarity-based. Only one event sample is produced by each of these methods, and hence no sub-components are shown.

4.3 Unitarity

The oldest, and in my view most attractive, approach [106, 136] consists of working out the shower approximation to a given fixed order, and correcting the shower splitting functions at that order by a multiplicative factor given by the ratio of the ME to the shower approximation, phase-space point by phase-space point. We may sketch this as

$$\text{Matched} = \overbrace{\text{Approximate}}^{\text{shower}} \times \overbrace{\frac{\text{Exact}}{\text{Approximate}}}^{\text{correction}} . \tag{79}$$

When these correction factors are inserted back into the shower evolution, they guarantee that the shower evolution off $n-1$ partons correctly reproduces the n-parton MEs, without the need to generate a separate n-parton sample. That is, the shower approximation is essentially used as a pre-weighted (stratified) all-orders phase-space generator, on which a more exact answer can subsequently be imprinted order by order in perturbation theory. Since the shower is already optimized for exactly the kind of singular structures that occur in QCD, very fast computational speeds can therefore be obtained with this method [81].

In the original approach [106, 136], used by PYTHIA [90, 91], this was only worked out for one additional emission beyond the basic hard process. In POWHEG [134, 137], it was extended to also include virtual corrections to the Born-level ME. Finally, in VINCIA, it has been extended to include arbitrary numbers of emissions at tree level [80, 81] and one emission at loop level [138], although that method has so far only been applied to final-state showers.

An illustration of the perturbative coefficients that can be included in each of these approaches is given in Fig. 27, as usual with green (darker shaded) boxes representing exact coefficients and yellow (light shaded) boxes representing logarithmic approximations.

Finally, two more properties unique to this method deserve a mention. Firstly, since the corrections modify the actual shower evolution kernels, the corrections are automatically *resummed* in the Sudakov exponential, which should improve the logarithmic precision once $k \geq 2$ is included, and secondly, since the shower is *unitary*, an initially unweighted sample of $(n-1)$-parton configurations remains unweighted, with no need for a separate event-unweighting or event-rejection step.

5 Hadronization and soft hadron–hadron physics

We here give a very brief overview of the main aspects of soft QCD that are relevant for hadron–hadron collisions, such as hadronization, minimum-bias and soft-inclusive physics, and the so-called underlying event. This will be kept at a pedestrian level and is largely based on the reviews in Refs. [5, 24, 139].

In the context of event generators, *hadronization* denotes the process by which a set of coloured partons (*after* showering) is transformed into a set of colour-singlet *primary* hadrons, which may then subsequently decay further. This non-perturbative transition takes place at the *hadronization scale* $Q_{\rm had}$, which by construction is identical to the IR cutoff of the parton shower. In the absence of a first-principles

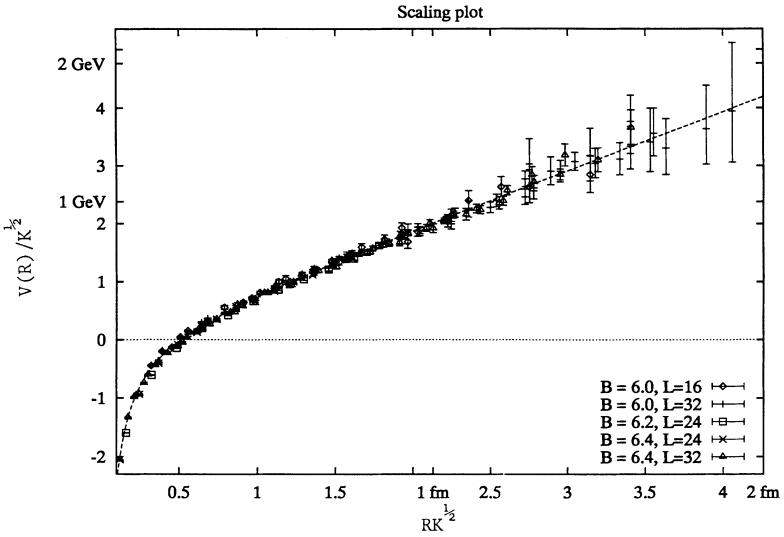


Fig. 28: Static quark–antiquark potential, as a function of separation distance, in quenched lattice QCD, from Ref. [140]. Note that the axes are scaled by units of the string tension $\sqrt{\kappa} \sim 420$ MeV. Additional labels corresponding to 1 GeV and 2 GeV are also provided, on the y-axis, and to 1 fm and 2 fm, on the x-axis. A constant term, V_0, has been subtracted from all the results. The dashed line corresponds to $V(R) = R - \pi/(12R)$.

solution to the relevant dynamics, event generators use QCD-inspired phenomenological models to describe this transition.

The problem can be stated as follows: given a set of partons resolved at a scale of $Q_{\rm had} \sim 1$ GeV, we need a 'mapping' from this set onto a set of on-shell colour-singlet (i.e., confined) hadronic states. MC models do this in three steps:

1. Map the partonic system onto a continuum of high-mass hadronic states (called 'strings' or 'clusters').
2. Iteratively map strings/clusters onto discrete set of primary hadrons (via string breaks / cluster splittings / cluster decays).
3. Sequential decays into secondaries ($\rho \to \pi\pi$, $\Lambda \to {\rm n}\pi$, $\pi^0 \to \gamma\gamma$, ...).

The physics governing this mapping is non-perturbative. However, we do have some knowledge of the properties that such a solution must have. For instance, Poincaré invariance, unitarity, and causality are all concepts that apply beyond perturbation theory. In addition, lattice QCD provides us a means of making explicit quantitative studies in a genuinely non-perturbative setting (albeit only of certain questions).

An important result in 'quenched' lattice QCD[24] is that the potential of the colour-dipole field between a charge and an anticharge appears to grow linearly with the separation of the charges, at distances greater than about 0.5 fm; this behaviour is illustrated by the plot shown in Fig. 28, from Ref. [140]. (Note that the axes are scaled by units of the string tension $\sqrt{\kappa} \sim 420$ MeV. Additional labels corresponding to 1 GeV and 2 GeV are also provided, on the y-axis, and to 1 fm and 2 fm, on the x-axis.) This is known as 'linear confinement', and it forms the starting point for the *string model of hadronization*, discussed below in Section 5.1. Alternatively, a property of pQCD called 'preconfinement' [141] is the basis of the *cluster model of hadronization*, described in Refs. [5, 24].

In the generator landscape, PYTHIA uses string fragmentation, while HERWIG and SHERPA use cluster fragmentation. It should be emphasized that the so-called *parton level* that can be obtained by

[24] Quenched QCD implies no 'dynamical' quarks, i.e., no ${\rm g} \to {\rm q\bar{q}}$ splittings allowed.

switching off hadronization in an MC generator, is not a universal concept, since each model defines the hadronization scale differently. For example, the hadronization scale can be defined by a cutoff in invariant mass, transverse momentum, or some other quantity, with different tunes using different values for the cutoff. Note that the so-called *parton level* that can be obtained by switching off hadronization in an MC generator, is not a universal concept, since each model defines the hadronization scale differently, with different tunes using different values for it. Comparisons to distributions at this level (i.e., with hadronization switched off) may therefore be used to provide an idea of the overall impact of hadronization corrections within a given model, but should be avoided in the context of physical observables. Note also that the corresponding MC *fragmentation functions* are intrinsically defined at the hadronization scale. They can therefore not be compared directly to those that are used in fixed-order/analytical-resummation contexts, which are typically defined at a factorization scale of the order of the scale of the hard process.

We use the term 'soft hadron–hadron physics' to comprise all scattering processes for which a hard, perturbative scale is not required to be present[25]. This includes elastic, diffractive, minimum-bias and pile-up processes, as well as the physics contributing to the so-called underlying event. We give a brief introduction to such processes in Section 5.2.

We round off with a discussion of the data constraints that enter in the tuning of Monte Carlo models in Section 5.4 and give an outline of a procedure that could be followed in a realistic set-up.

5.1 String model

Starting from early concepts developed by Artru and Mennessier [143], several hadronization models based on strings were proposed in the late 1970s and early 1980s. Of these, the most widely used today is the so-called Lund model, implemented in the PYTHIA code. We shall, therefore, concentrate on that particular model here, although many of the overall concepts would be shared by any string-inspired method. (A more extended discussion can be found in the very complete and pedagogical review of the Lund model by Andersson [6].)

Consider the production of a $\mathrm{q\bar{q}}$ pair from vacuum, for instance in the process $\mathrm{e^+e^-} \to \gamma^*/\mathrm{Z} \to \mathrm{q\bar{q}} \to$ hadrons. As the quarks move apart, linear confinement implies that a potential

$$V(R) = \kappa\, R \tag{80}$$

is asymptotically reached for large distances, R. At short distances, there is a Coulomb term proportional to $1/R$ as well, cf. Fig. 28, but this is neglected in the Lund model. Such a potential describes a string with tension (energy per unit length) κ, with the value [140]

$$\kappa \;\sim\; (420\,\mathrm{MeV})^2 \;\sim\; 0.18\,\mathrm{GeV}^2 \;\sim\; 0.9\,\mathrm{GeV/fm}, \tag{81}$$

which, for comparison with the world of macroscopic objects, would be sufficient to lift a 16-ton truck [144].

The string can be thought of as parameterizing the position of the axis of a cylindrically symmetric flux tube, illustrated in Fig. 29. Such simple $\mathrm{q} - \bar{\mathrm{q}}$ strings form the starting point for the string model. More complicated multi-parton topologies are treated by representing gluons as transverse 'kinks', e.g., $\mathrm{q} - \mathrm{g} - \bar{\mathrm{q}}$. The space–time evolution is then slightly more involved [6], and modifications to the fragmentation model to handle stepping across gluon corners have to be included, but the main point is that there are no separate free parameters for gluon jets. Differences with respect to quark fragmentation arise simply because quarks are only connected to a single string piece, while gluons have one on either side,

[25] Note, however, that while a hard scale is not *required* to be present, it is not explicitly required to be absent either. Thus, both diffractive, minimum-bias, pile-up and underlying-event processes will have tails towards high-$p_\perp$ physics as well. For example, even $t\bar{t}$ pair production can be viewed as a tail of minimum-bias interactions, and there is a tail of diffractive processes in which hard dijets can be produced diffractively (see, e.g., Ref. [142]).

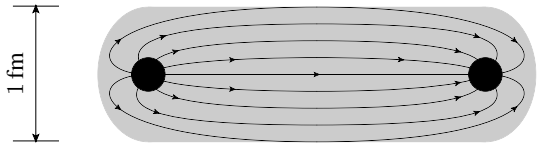


Fig. 29: Illustration of the transition between a Coulomb potential at short distances to the string-like one of Eq. (80) at large $\mathrm{q\bar{q}}$ separations.

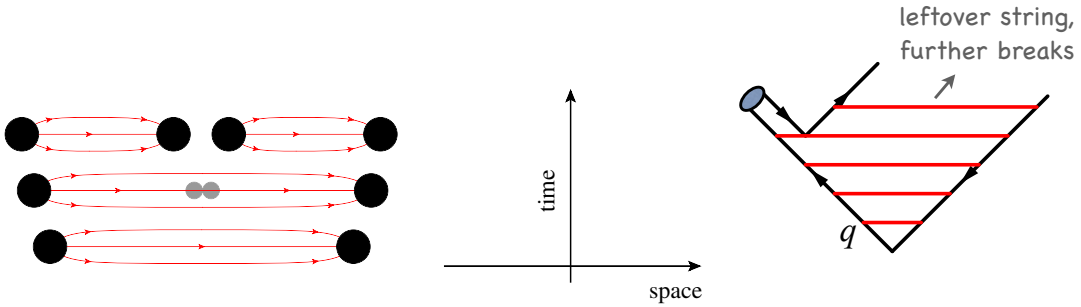


Fig. 30: (a) Illustration of string breaking by quark pair creation in the string field. (b) Illustration of the algorithmic choice to process the fragmentation from the outside-in, splitting off a single on-shell hadron in each step.

increasing the energy loss per unit (invariant) time from a gluon to the string by a factor of two relative to quarks, which can be compared to the ratio of colour Casimirs $C_A/C_F = 2.25$. Another appealing feature of the model is that low-energy gluons are absorbed smoothly into the string, without leading to modifications. This improves the stability of the model with respect to variations of the IR behaviour of the parton shower.

As the partonic string endpoints move apart, their kinetic energy is gradually converted to potential energy, stored in the growing string spanned between them. In the 'quenched' approximation, in which $\mathrm{g} \to \mathrm{q\bar{q}}$ splittings are not allowed, this process would continue until the endpoint quarks have lost *all* their momentum, at which point they would reverse direction and be accelerated by the now shrinking string.

In the real world, quark–antiquark fluctuations inside the string field can make the transition to become real particles by absorbing energy from the string, thereby screening the original endpoint charges from each other and breaking the string into two separate colour-singlet pieces, $(\mathrm{q\bar{q}}) \to (\mathrm{q\bar{q}}') + (\mathrm{q}'\bar{\mathrm{q}})$, illustrated in Fig. 30 (a). This process then continues until only ordinary hadrons remain. (We will give more details on the individual string breaks below.)

Since the string breaks are causally disconnected (as can easily be realized from space–time diagrams like the one in Fig. 30, see also Ref. [6]), they do not have to be considered in any specific time-ordered sequence. In the Lund model, the string breaks are instead generated starting with the leading ('outermost') hadrons, containing the endpoint quarks, and iterating inwards towards the centre of the string, alternating randomly between fragmentation off the left- and right-hand sides, respectively, Fig. 30 (b). One can thereby split off a single well-defined hadron in each step, with a mass that, for unstable hadrons, is selected according to a Breit–Wigner distribution.

The details of the individual string breaks are not known from first principles. The Lund model invokes the idea of quantum mechanical tunneling, which leads to a Gaussian suppression of the transverse momenta and masses imparted to the produced quarks,

$$\mathrm{Prob}(m_{\mathrm{q}}^2, p_{\perp q}^2) \;\propto\; \exp\left(\frac{-\pi m_{\mathrm{q}}^2}{\kappa}\right) \exp\left(\frac{-\pi p_{\perp \mathrm{q}}^2}{\kappa}\right), \tag{82}$$

where m_{q} is the mass of the produced quark and $p_\perp$ is the transverse momentum imparted to it by the breakup process (with the $\bar{\mathrm{q}}$ having the opposite $p_\perp$).

Due to the factorization of the $p_\perp$and m dependence implied by Eq. (82), the $p_\perp$spectrum of produced quarks in this model is independent of the quark flavour, with a universal average value of

$$\left\langle p_{\perp q}^2 \right\rangle = \sigma^2 = \kappa/\pi \sim (240\,\mathrm{MeV})^2. \tag{83}$$

Bear in mind that 'transverse' is here defined with respect to the string axis. Thus, the $p_\perp$in a frame where the string is moving is modified by a Lorentz boost factor. Also bear in mind that σ^2 is here a purely non-perturbative parameter. In a Monte Carlo model with a fixed shower cutoff Q_{had}, the effective amount of 'non-perturbative' $p_\perp$may be larger than this, due to effects of additional unresolved soft-gluon radiation below Q_{had}. In principle, the magnitude of this additional component should scale with the cutoff, but in practice it is up to the user to enforce this by retuning (see Section 5.4) the effective σ parameter when changing the hadronization scale. Since hadrons receive $p_\perp$ contributions from two breakups, one on either side, their average transverse momentum squared will be twice as large,

$$\left\langle p_{\perp \mathrm{h}}^2 \right\rangle = 2\sigma^2. \tag{84}$$

The mass suppression implied by Eq. (82) is less straightforward to interpret. Since quark masses are notoriously difficult to define for light quarks, the value of the strangeness suppression must effectively be extracted from experimental measurements, for example, of the K/π ratio, with a resulting suppression of roughly $s/u \sim s/d \sim 0.2$–0.3. Inserting even comparatively low values for the charm quark mass in Eq. (82), however, one obtains a relative suppression of charm of the order of 10^{-11}. Heavy quarks can therefore safely be considered to be produced only in the perturbative stages and not by the soft fragmentation.

Baryon production can be incorporated in the same basic picture [145], by allowing string breaks to occur also by the production of pairs of so-called *diquarks*, loosely bound states of two quarks in an overall $\bar{3}$ representation (e.g., 'red + blue $\sim$ antigreen', cf. the rules for colour combinations in Section 1.2). Again, the relative rate of diquark-to-quark production is not known a priori and must be extracted from experimental measurements, for example, of the p/π ratio. More advanced scenarios for baryon production have also been proposed, in particular the so-called popcorn model [146, 147], which is normally used in addition to the diquark picture and then acts to decrease the correlations among neighbouring baryon–antibaryon pairs by allowing mesons to be formed inbetween them. Within the PYTHIA framework, a fragmentation model including explicit *string junctions* [148] has so far only been applied to baryon-number-violating new-physics processes and to the description of beam remnants (and then acts to increase baryon stopping [149]).

This brings us to the next step of the algorithm: assignment of the produced quarks within hadron multiplets. Using a nonrelativistic classification of spin states, the fragmenting q ($\bar{\mathrm{q}}$) may combine with the $\bar{\mathrm{q}}'$ (q') from a newly created breakup to produce either a vector or a pseudoscalar meson, or, if diquarks are involved, either a spin-$1/2$ or spin-$3/2$ baryon. Unfortunately, the string model is entirely unpredictive in this respect, and this is therefore the sector that contains the largest amount of free parameters. From spin counting alone, one would expect the ratio V/S of vectors to pseudoscalars to be three, but this is modified by the V–S mass splittings, which implies a phase-space suppression of vector production, with corresponding suppression parameters to be extracted from data.

Especially for the light flavours, the substantial difference in phase space caused by the V–S mass splittings implies a rather large suppression of vector production. Thus, for D^*/D, the effective ratio is already reduced to about $\sim$ 1.0–2.0, while for K^*/K and ρ/π, extracted values range from 0.3–1.0. (Recall, as always, that these are production ratios of *primary hadrons*, hence feed-down from secondary decays of heavier hadrons complicates the extraction of these parameters from experimental data, in particular for the lighter hadron species.)

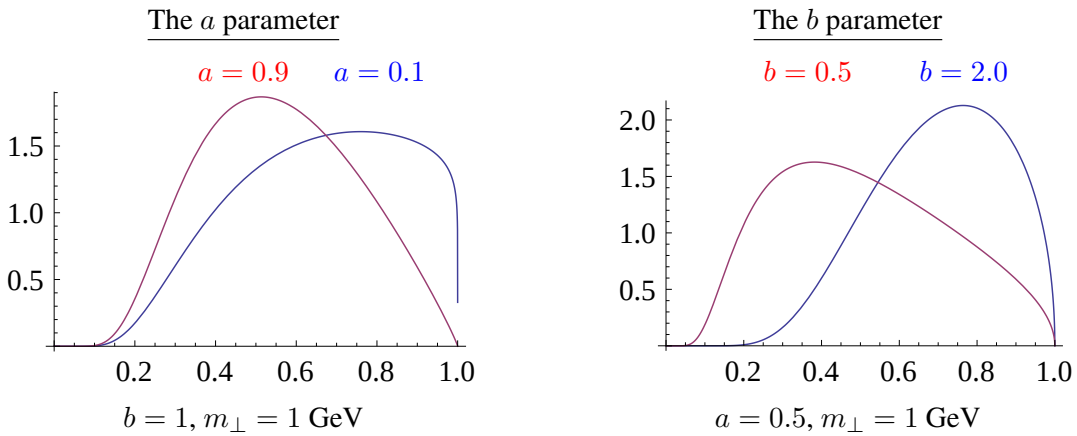


Fig. 31: Normalized Lund symmetric fragmentation function, for fixed $m_\perp = 1$ GeV. *Left:* variation of the a parameter, from 0.1 (blue) to 0.9 (red), with fixed $b = 1$ GeV^{-2}. *Right:* variation of the b parameter, from 0.5 (red) to 2 (blue) GeV^{-2}, with fixed $a = 0.5$.

The production of higher meson resonances is assumed to be low in a string framework[26]. For diquarks, separate parameters control the relative rates of spin-1 diquarks versus spin-0 ones and, likewise, have to extracted from data, with resulting values of order $(qq)_1/(qq)_0 \sim 0.075$–0.15.

With $p_\perp^2$ and m^2 now fixed, the final step is to select the fraction, z, of the fragmenting endpoint quark's longitudinal momentum that is carried by the created hadron. In this respect, the string picture is substantially more predictive than for the flavour selection. Firstly, the requirement that the fragmentation be independent of the sequence in which breakups are considered (causality) imposes a 'left-right symmetry' on the possible form of the fragmentation function, $f(z)$, with the solution [150]

$$f(z) \propto \frac{1}{z}(1-z)^a \exp\left(-\frac{b\,(m_{\mathrm{h}}^2 + p_{\perp\mathrm{h}}^2)}{z}\right), \tag{85}$$

which is known as the *Lund symmetric fragmentation function* (normalized to unit integral). The a and b parameters, illustrated in Fig. 31, are the only free parameters of the fragmentation function, although a may in principle be flavour-dependent. Note that the explicit mass dependence in $f(z)$ implies a harder fragmentation function for heavier hadrons (in the rest frame of the string).

For massive endpoints (e.g., c and b quarks), which do not move along straight lightcone sections, the exponential suppression with string area leads to modifications of the form [151], $f(z) \to f(z)/z^{b\,m_{\mathrm{Q}}^2}$, with m_{Q} the heavy-quark mass. Strictly speaking, this is the only fragmentation function that is consistent with causality in the string model, although a few alternative forms are typically provided as well.

As a by-product, the probability distribution in invariant time τ of $\mathrm{q}'\bar{\mathrm{q}}'$ breakup vertices, or equivalently $\Gamma = (\kappa\tau)^2$, is also obtained, with $\mathrm{d}P/\mathrm{d}\Gamma \propto \Gamma^a \exp(-b\Gamma)$ implying an area law for the colour flux [152], and the average breakup time lying along a hyperbola of constant invariant time $\tau_0 \sim 10^{-23}$ s [6].

We may also ask, for example, how many units of rapidity does the particle production from a string span? Measuring p_z along the string direction and defining rapidity by

$$y = \frac{1}{2}\ln\left(\frac{E+p_z}{E-p_z}\right), \tag{86}$$

[26]The four $L = 1$ multiplets are implemented in PYTHIA, but are disabled by default, largely because several states are poorly known and thus may result in a worse overall description when included.

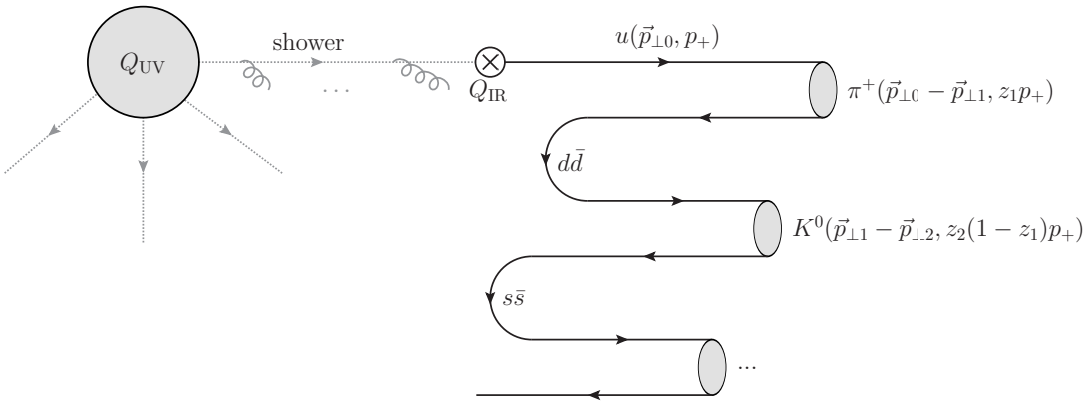


Fig. 32: Iterative selection of flavours and momenta in the Lund string-fragmentation model

the absolute highest rapidity that can be reached, by a pion traveling exactly along the string direction and taking all of the endpoint quark's energy, is $y_{\max} = \ln(2E_{\mathrm{q}}/m_\pi)$. That is, the rapidity region covered by a fragmenting string scales logarithmically with the energy, and since the density of hadrons produced per unit rapidity is roughly constant (modulo endpoint effects), the average number of hadrons produced by string fragmentation likewise scales logarithmically with energy.

The iterative selection of flavours, $p_\perp$, and z values is illustrated in Fig. 32. A parton produced in a hard process at some high scale Q_{UV} emerges from the parton shower, at the hadronization scale Q_{IR}, with 3-momentum $\vec{p} = (\vec{p}_{\perp 0}, p_+)$, where the '+' on the third component denotes 'light-cone' momentum, $p_\pm = E \pm p_z$. Next, an adjacent $\mathrm{d\bar{d}}$ pair from the vacuum is created, with relative transverse momenta $\pm p_{\perp 1}$. The fragmenting quark combines with the $\bar{\mathrm{d}}$ from the breakup to form a π^+, which carries off a fraction z_1 of the total lightcone momentum p_+. The next hadron carries off a fraction z_2 of the remaining momentum, etc.

5.2 Soft hadron–hadron processes

The total hadron–hadron (hh) cross section is around 100 mb at LHC energies [153], growing slowly with the CM energy, $\sigma_{\mathrm{tot}}(s) \propto s^{0.096}$ [154]. There are essentially four types of physics processes, which together make up σ_{tot}:

1. elastic scattering: hh $\to$ hh;
2. single diffractive dissociation: $\mathrm{hh} \to \mathrm{h} + \mathrm{gap} + \mathrm{X}$, with 'gap' denoting an empty rapidity region, and X anything that is not the original beam particle;
3. double diffractive dissociation: $\mathrm{hh} \to \mathrm{X} + \mathrm{gap} + \mathrm{X}$ (both hadrons 'blow up');
4. Inelastic non-diffractive scattering: everything else.

In principle, higher 'multi-gap' diffractive components may be defined as well, the most important one being central diffraction (CD): $\mathrm{hh} \to \mathrm{h} + \mathrm{gap} + \mathrm{X} + \mathrm{gap} + \mathrm{h}$, see the discussion of diffraction in Section 5.2.1 below.

Some important differences exist between theoretical and experimental terminology [155]. In the experimental setting, diffraction is defined by an observable rapidity gap, with $|\Delta y|_{\mathrm{gap}} \gtrsim 3$ typically giving clean diffractive samples. In the MC context, however, each diffractive process type produces a whole spectrum of gaps, with small ones suppressed but not excluded. Likewise, events of non-diffractive origin may produce accidental rapidity gaps, now with large ones suppressed (exponentially) but not excluded, and in the transition region there could even be quantum mechanical interference between the two. Due to this unphysical model dependence of theoretical definitions of diffraction, we strongly

advise to phrase measurements first and foremost in terms of physical observables, and only seek to connect with theory models as a second, separate, step.

The distinction between elastic and inelastic events *is*, however, unambiguous (modulo $\mathrm{pp} \to \mathrm{pp}\gamma$ events); the final state either contains just two protons, or not. The total hadron–hadron cross section can therefore be written as a sum of these two physically distinguishable components,

$$\sigma_{\mathrm{tot}}(s) = \sigma_{\mathrm{el}}(s) + \sigma_{\mathrm{inel}}(s) \,, \tag{87}$$

where $s = (p_A + p_B)^2$ is the beam–beam centre-of-mass energy squared.

Another potentially confusing term is 'minimum bias' (MB). This originates from the experimental requirement of a minimal energy or number of hits in a given (experiment-dependent) instrumented region near the beam, used to determine whether there was any non-trivial activity in the event, or not. This represents the smallest possible 'trigger bias' that the corresponding experiment is capable of. There is thus no universal definition of 'min-bias'; each experiment has its own. We give a brief discussion of MB in Section 5.2.2 below.

Finally, in events containing a hard parton–parton interaction, the UE can be roughly conceived of as the *difference* between QCD with and without including the remnants of the original beam hadrons. Without such 'beam remnants', only the hard interaction itself, and its associated parton showers and hadronization, would contribute to the observed particle production. In reality, after the partons that participate in the hard interaction have been taken out, the remnants still contain whatever is left of the incoming beam hadrons, including also a partonic substructure, which leads to the possibility of *multiple parton interactions* (MPI). We discuss MPI-based models of the UE in Section 5.3 below. Other useful reviews of MPI-based MC models can be found in Refs. [5, 139]. Analytical models are mostly formulated only for double parton scattering, see, for example, Refs. [156–159].

5.2.1 Diffractive scattering

As mentioned above, if the beam particles A and/or B are not elementary, the inelastic final states may be divided into 'diffractive' and 'non-diffractive' topologies. This is a qualitative classification, usually based on whether the final state looks like the decay of an excitation of the beam particles (diffractive[27]), or not (non-diffractive), or upon the presence of a large rapidity gap somewhere in the final state which would separate such excitations.

Given that an event has been labeled as diffractive, either within the context of a theoretical model, or by a final-state observable, we may distinguish between three different classes of diffractive topologies, which it is possible to distinguish between physically, at least in principle. In double-diffractive (DD) events, both of the beam particles are diffractively excited and hence none of them survive the collision intact. In single-diffractive (SD) events, only one of the beam particles gets excited and the other survives intact. The last diffractive topology is CD, in which both of the beam particles survive intact, leaving an excited system in the central region between them. (This latter topology includes 'central exclusive production' where a single particle is produced in the central region.) That is,

$$\sigma_{\mathrm{inel}}(s) = \sigma_{\mathrm{SD}}(s) + \sigma_{\mathrm{DD}}(s) + \sigma_{\mathrm{CD}}(s) + \sigma_{\mathrm{ND}}(s) \,, \tag{88}$$

where 'ND' (non-diffractive, here understood not to include elastic scattering) contains no gaps in the event consistent with the chosen definition of diffraction. Further, each of the diffractively excited systems in the events labeled SD, DD and CD, respectively, may in principle consist of several subsystems with gaps between them. Eq. (88) may thus be defined to be exact, within a specific definition of diffraction, even in the presence of multi-gap events. Note, however, that different theoretical models almost

[27] An example of a process that would be labeled as diffractive would be if one the protons is excited to a Δ^+ which then decays back to $\mathrm{p}^+ + \pi^0$, without anything else happening in the event. In general, a whole tower of possible diffractive excitations are available, which in the continuum limit can be described by a mass spectrum falling roughly as $\mathrm{d}M^2\,/M^2$.

always use different (model-dependent) definitions of diffraction, and therefore the individual components in one model are in general not directly comparable to those of another. It is therefore important that data be presented at the level of physical observables if unambiguous conclusions are to be drawn from them.

5.2.2 Minimum bias

In principle, *everything* that produces a hit in a given experiment's minimum-bias trigger, is a subset of MB. In particular, since there is no explicit veto on hard activity, it is useful to keep in mind that MB includes a diverse mixture of both soft and hard processes, although the fraction that is made up of hard high-$p_\perp$ processes is only a small tail compared to the total minimum-bias cross section[28].

In theoretical contexts, the term 'minimum-bias' is often used with a slightly different meaning; to denote specific (classes of) inclusive soft-QCD subprocesses in a given model. Since these two usages are not exactly identical, in these lectures we have chosen to reserve the term 'MB' to pertain strictly to definitions of experimental measurements, and instead use the term 'soft inclusive' physics as a generic descriptor for the class of processes which generally dominate the various experimental 'MB' measurements in theoretical models. This parallels the terminology used in the review Ref. [5], from which most of the discussion here has been adapted. See Eq. (88) above for a compact overview of the types of physical processes that contribute to minimum-bias data samples. For a more detailed description of Monte Carlo models of this physics, see Ref. [5].

5.3 Underlying event and multiple parton interactions

In this subsection, we focus on the physics of MPI as a theoretical basis for understanding both inelastic, non-diffractive processes (MB), as well as the so-called underlying event (a.k.a. the jet pedestal effect). Keep in mind, however, that especially at low multiplicities, and when gaps are present, the contributions from diffractive processes should not be ignored.

Due to the simple fact that hadrons are composite, multi-parton interactions (several distinct parton–parton interactions in one and the same hadron–hadron collision) will always be there—but how many, and how much additional energy and tracks do they deposit in a given measurement region? The first detailed Monte Carlo model for perturbative MPI was proposed by Sjöstrand in Ref. [160], and with some variation this still forms the basis for most modern implementations [5].

The first crucial observation is that the t-channel propagators appearing in pQCD $2 \to 2$ scattering almost go on shell at low $p_\perp$, causing the differential cross sections to become very large, behaving roughly as

$$d\sigma_{2\to 2} \propto \frac{dt}{t^2} \sim \frac{dp_\perp^2}{p_\perp^4} \,. \tag{89}$$

At LHC energies, this *parton–parton* cross section becomes larger than the total *hadron–hadron* cross section at $p_\perp$ scales of order 4–5 GeV. This is illustrated in Fig. 33, in which we compare the integrated QCD parton–parton cross section (with two different α_s and PDF choices) to the total inelastic cross section measured by TOTEM [153], for pp collisions at $\sqrt{s} = 8$ TeV. In the context of MPI models, this is interpreted straightforwardly to mean that *each* hadron–hadron collision contains *several* few GeV parton–parton collisions.

In the limit that all the partonic interactions are independent and equivalent, one would simply have a Poisson distribution in the number of MPI, with average

$$\langle n \rangle (p_{\perp\mathrm{min}}) = \frac{\sigma_{2\to 2}(p_{\perp\mathrm{min}})}{\sigma_\mathrm{tot}}, \tag{90}$$

[28] Furthermore, since only a tiny fraction of the total minimum-bias rate can normally be stored, the minimum-bias sample would give quite poor statistics if used for hard physics studies. Instead, separate dedicated hard-process triggers are typically included in addition to the minimum-bias one, in order to ensure maximal statistics also for hard physics processes.

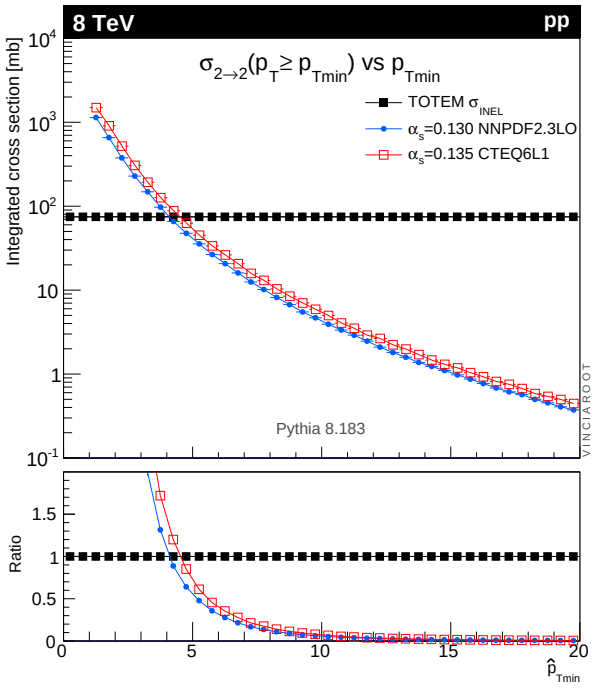


Fig. 33: Proton-proton collisions at 8 TeV. LO QCD parton–parton cross section (integrated above $p_{T\min}$, for two different α_s and PDF choices) compared to the total inelastic hadron–hadron cross section. Towards the right of the plot, we see, as expected, that hard dijet events is only a tiny fraction of the total cross section. The fact that the curves cross at a scale of order 5 GeV is interpreted to mean that this is a characteristic scale relevant for MPI. [161].

with $p_{\perp\min}$ a lower cutoff scale which we shall return to below, and $\sigma_{\rm tot}$ a measure of the inelastic hadron–hadron cross section. This simple reinterpretation in fact expresses unitarity; instead of the total interaction cross section diverging as $p_{\perp\min} \to 0$ (which would violate unitarity), we have restated the problem so that it is now the *number of MPI per collision* that diverges, with the total cross section remaining finite.

Two important ingredients remain to fully regulate the remaining divergence. Firstly, the interactions cannot use up more momentum than is available in the parent hadron. This suppresses the large-n tail of the estimate above. In PYTHIA-based models, the MPI are ordered in $p_\perp$ [111, 160], and the parton densities for each successive interaction are explicitly constructed so that the sum of x fractions can never be greater than unity. In the HERWIG models [162, 163], instead the uncorrelated estimate of $\langle n \rangle$ above is used as an initial guess, but the generation of actual MPI is stopped once the energy-momentum conservation limit is reached.

The second ingredient invoked to suppress the number of interactions, at low $p_\perp$ and x, is colour screening; if the wavelength $\sim 1/p_\perp$ of an exchanged coloured parton becomes larger than a typical colour–anticolour separation distance, it will only see an *average* colour charge that vanishes in the limit $p_\perp \to 0$, hence leading to suppressed interactions. This provides an IR cutoff for MPI similar to that provided by the hadronization scale for parton showers. A first estimate of the colour-screening cutoff would be the proton size, $p_{\perp\min} \approx \hbar/r_p \approx 0.3$ GeV $\approx \Lambda_{\rm QCD}$, but empirically this appears to be far too low. In current models, one replaces the proton radius $r_{\rm p}$ in the above formula by a 'typical colour screening distance', that is, an average size of a region within which the net compensation of a given colour charge occurs. This number is not known from first principles, although it may be related to saturation [164]. In current MPI models, it is perceived of simply as an effective cutoff parameter, to be determined from data.

Note that the partonic cross sections depend upon the PDF set used, and therefore the optimal value to use for the cutoff will also depend on this choice [165]. Note also that the cutoff does not have to be energy-independent. Higher energies imply that parton densities can be probed at smaller x values, where the number of partons rapidly increases. Partons then become closer packed and the colour-screening distance d decreases. The uncertainty on the scaling of the cutoff is a major concern when extrapolating between different collider energies [165–167].

We now turn to the origin of the observational fact that hard jets appear to sit on top of a higher 'pedestal' of underlying activity than events with no hard jets. That is, the so-called UE is much more active, with larger fluctuations, than the average min-bias event. This is interpreted as a consequence of impact-parameter-dependence: in peripheral collisions, only a small fraction of events contain any high-$p_\perp$ activity, whereas central collisions are more likely to contain at least one hard scattering; a high-$p_\perp$ triggered sample will therefore be biased towards small impact parameters, b, with a large number of MPI (and associated increased activity). The ability of a model to describe the shape of the pedestal (e.g., to describe both MB and UE distributions simultaneously) therefore depends upon its modelling of the b-dependence, and correspondingly the impact-parameter shape constitutes another main tuning parameter. A detailed discussion of impact-parameter dependent models goes beyond the scope of these lectures; see Refs. [149, 160, 168].

For hard processes at the LHC at 13 TeV, the transverse energy, E_T, in the UE is expected to be about 3.3 GeV per unit $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ area [167] (for a reference case of 100 GeV dijets, including both charged and neutral particles, with no cut on $p_\perp$), although with large event-to-event fluctuations of order ± 2 GeV [169]. Thus, for example, the total E_{T} originating from the UE, in a cone with radius 0.4 can be estimated to be $E_{T\mathrm{UE}}(R = 0.4) \sim 1.6 \pm 1$ GeV, while the E_T in a cone with radius 1.0 would be $E_{T\mathrm{UE}}(R = 1.0) \sim 10 \pm 6$ GeV. Note that the magnetic field in realistic detectors will deflect some fraction of the soft charged component of the underlying event into helix trajectories that will hence not contribute to the energy deposition in the calorimeters.

5.4 Tuning

A main virtue of general-purpose Monte Carlo event generators is their ability to provide a complete and fully differential picture of collider final states, down to the level of individual particles. As has been emphasized in these lectures, the achievable accuracy depends both on the inclusiveness of the chosen observable and on the sophistication of the simulation itself. An important driver for the latter is obviously the development of improved theoretical models, for example, by including matching to higher-order MEs, more accurate resummations or better non-perturbative models, as discussed in the previous sections, but it also depends crucially on the available constraints on the remaining free parameters of the model. Using existing data (or more precise calculations) to constrain these is referred to as generator tuning.

Keep in mind that generators attempt to deliver a *global* description of the data; a tune is no good if it fits one distribution perfectly, but not any others. It is therefore crucial to study the simultaneous degree of agreement or disagreement over many, mutually complementary, distributions. A useful online resource for making such comparisons can be found at the MCPLOTS web site [170] (which relies on computing power donated by volunteers, via the LHC@home project [171]). The analyses come from the comprehensive RIVET analysis toolkit [95], which can also be run stand-alone to make your own MC tests and comparisons.

Although MC models may appear to have a bewildering number of independently adjustable parameters, it is worth noting that most of these only control relatively small (exclusive) details of the event generation. The majority of the (inclusive) physics is determined by only a few, very important ones, such as the value of the strong coupling, in the perturbative domain, and the form of the fragmentation function for massless partons, in the non-perturbative one.

Armed with a good understanding of the underlying model, an expert would therefore normally take a highly factorized approach to constraining the parameters, first constraining the perturbative ones (using IR-safe observables and/or more precise theory calculations) and thereafter the non-perturbative ones, each ordered in a measure of their relative significance to the overall modelling. This allows one to concentrate on just a few parameters and a few carefully chosen distributions at a time, reducing the full parameter space to manageable-sized chunks. Still, each step will often involve more than one single parameter, and non-factorizable correlations may still necessitate additional iterations from the beginning before a fully satisfactory set of parameters is obtained.

Recent years have seen the emergence of automated tools that attempt to reduce the amount of both computer and manpower required for this task, for instance by making full generator runs only for a limited set of parameter points, and then interpolating between these to obtain approximations to what the true generator result would have been for any intermediate parameter point, as, for example, in PROFESSOR [172]. Automating the human expert input is more difficult. Currently, this is addressed by a combination of input solicited from the generator authors (e.g., which parameters and ranges to consider, which observables constitute a complete set, etc) and the elaborate construction of non-trivial weighting functions that determine how much weight is assigned to each individual bin in each distribution. The field is still burgeoning, and future sophistications are to be expected. Nevertheless, at this point the overall quality of the tunes obtained with automated methods appear to at least be competitive with the manual ones.

However, although we have very good LHC tunes for essentially all the general-purpose generators by now, there are two important aspects which have so far been neglected, and which it is becoming increasingly urgent to address. The first is that a central tune is not really worth much, unless you know what the uncertainty on it is. A few individual proposals for systematic tuning variations have been made [166, 173], but so far there is no general approach for establishing MC uncertainties by tune variations. The second issue is that virtually all generator tuning is done at the 'pure' LL-shower level, and not much is known about what happens to the tuning when matrix-element matching is subsequently included.

Finally, rather than performing one global tune to all the data, as is usually done, a more systematic check on the validity of the underlying physics model could be obtained by instead performing several independent optimizations of the model parameters for a range of different phase-space windows and/or collider environments. In regions in which consistent parameter sets are obtained (with reasonable $\Delta\chi^2$ values), the underlying model can be considered as interpolating well, that is, it is universal. If not, a breakdown in the ability of the model to span different physical regimes has been identified, and can be addressed, with the nature of the deviations giving clues as to the nature of the breakdown. With the advent of automated tools, such systematic studies are now becoming feasible, with a first example given in Ref. [165].

We round off by giving a sketch of a reasonably complete tuning procedure, without going into details about the parameters that control each of these sectors in individual Monte Carlo models (a recent detailed discussion in the context of PYTHIA 8 can be found in Ref. [161]).

1) **Keep in mind** that inabilities of models to describe data is a vital part of the feedback cycle between theory and experiment. Also keep in mind that perturbation theory at (N)LO+LL is doing *very well* if it gets within 10% of a given IR-safe measurement. An agreement of 5% should be considered the absolute sanity limit, beyond which it does not make any sense to tune further. For some quantities, for example, ones for which the underlying modelling is *known* to be poor, an order-of-magnitude agreement or worse may have to be accepted.

2) **Final-state radiation and hadronization:** mainly using LEP and other $\mathrm{e^+e^-}$ collider data. On the IR safe side, there are event shapes and jet observables. On the IR sensitive side, multiplicities and particle spectra. Pay attention to the high-z tail of the fragmentation, where a single hadron carries a large fraction of an entire jet's momentum (most likely to give 'fakes'). Depending on the focus of the tuning,

attention should also be paid to identified-particle rates and ratios (perhaps with a nod to hadron-collider measurements), and to fragmentation in events containing heavy quarks and/or gluon jets. Usually, more weight is given to those particles that are most copiously produced. The scaling properties of IR-safe versus IR-sensitive contributions can be tested by comparing data at several different $\mathrm{e^+e^-}$ collider energies.

3) **Initial-state radiation, and 'primordial**[29] k_T**':** the main constraining distribution is the dilepton $p_\perp$ distribution in Drell–Yan events in hadron–hadron collisions. Ideally, one would like to use several different Q^2 values, and/or complementary processes, like $p_\perp(\mathrm{dijet})$ or $p_\perp(\mathrm{t\bar{t}})$. For any observables containing explicit jets, be aware that the UE can produce small horizontal shifts in jet $p_\perp$ distributions, which may in turn result in larger-than-expected vertical changes if the distributions are falling sharply. Also note that the ISR evolution is sensitive to the choice of PDFs.

4) **Initial–final connections:** radiation from colour lines connecting the initial and final states. DIS and jet broadening in hadron collisions. This is one of the most poorly controlled parts of most MC models, highly sensitive to the treatment of coherence (see, e.g., Ref. [174] for illustrations). Keep in mind that it is *not* directly constrained by pure final-state observables, such as LEP fragmentation, nor by pure initial-state observables, such as the Drell–Yan $p_\perp$ spectrum, which is why we list it as a separate item here. The modelling of this aspect can have important effects on specific observables, a recent example being the $\mathrm{t\bar{t}}$ forward-backward asymmetry at the Tevatron [175].

5) **Underlying event:** good constraints on the overall level of the underlying event can be obtained by counting the summed transverse energy (more IR safe) and/or particle multiplicities and average transverse momenta (more IR sensitive) in regions *transverse* to a hard trigger jet (more IR safe) or particle (more IR sensitive), see, for example, Ref. [176]. Constraints on the *fluctuations* of the underlying event are also important, and can be obtained, for example, by comparing to measurements of the RMS of such distributions [169]. Again, note that the UE is sensitive to the choice of PDFs [165].

6) **Colour (re-)connections and other final-state interactions:** by final-state interactions, we intend a broad spectrum of possible collective effects that may be included to a greater or lesser extent in various models. These effects include: Bose–Einstein correlations (see, e.g., Ref. [177]), rescattering (see, e.g., Ref. [178]), colour reconnections / string interactions (see, e.g., Ref. [179–181]), hydrodynamics (see, e.g., Ref. [182]), etc. As a rule, these effects are soft and/or non-perturbative and hence should not modify hard IR-safe observables appreciably. They can, however, have *drastic* effects on IR-sensitive ones, such as particle multiplicities, momentum distributions, and correlations, wherefore useful constraints are typically furnished by measurements of spectra and correlations as functions of quantities believed to serve as indicators of the strength of these phenomena (such as event multiplicity), and/or by collective-flow-type measurements. Finally, if the model includes a universal description of underlying event and soft-inclusive QCD, as many MPI-based models do, then minimum-bias data can also be used as a control sample, although one must then be careful either to address diffractive contributions properly or to include only gap-suppressed data samples. A complete MB and UE model should also be able to describe the rise of the pedestal from MB to UE, for example, in the transverse UE observables (see above).

7) **Beam remnants:** constraints on beam remnant fragmentation (see, e.g., Ref. [149]) are most easily obtained in the forward region, but,for example, the amount of baryon transport from the remnant to a given rapidity region can also be used to probe how much the colour structure of the remnant was effectively disturbed, with more baryon transport indicating a larger amount of 'beam baryon blowup'.

[29] Primordial k_T: an additional soft $p_\perp$ component that is injected on top of the $p_\perp$ generated by the initial-state shower itself, see Ref. [5, Section 7.1].

Acknowledgments

Thanks go to C. Brust, M. Cacciari, Y. Dokshitzer, L. Hartgring, A. Kronfeld, A. Larkoski, J. Lopez-Villarejo, C. Murphy, P. Nason, P. Pigard, S. Prestel, G. Salam, and T. Sjöstrand whose valuable comments and sharing of insight contributed to these lectures. In addition, I have used material from my 2010 ESHEP lectures [139] and 2014 AEPSHEP lectures, and from the ESHEP lectures by Mangano [183], by Salam [4, 184], by Sjöstrand [41], and by Stirling [185], as well as the recent review on Monte Carlo event generators by the MCnet collaboration [5].

References

[1] M.E. Peskin and D.V. Schroeder, *An Introduction to Quantum Field Theory* (Addison-Wesley, 1995).

[2] R.K. Ellis, W.J. Stirling and B.R. Webber, *Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol.* **8** (1996) 1.

[3] G. Dissertori, I.G. Knowles and M. Schmelling, *Quantum Chromodynamics — High Energy Experiments and Theory* (Oxford Science Publications, Oxford, 2003).

[4] Gavin P. Salam, Elements of QCD for hadron colliders, Proc. 17th European School on High-Energy Physics (ESHEP), Bautzen, Germany, 2009, arXiv:1012.4643.

[5] A. Buckley *et al.*, *Phys. Rept.* **504**(5) (2011) 145. http://dx.doi.org/10.1016/j.physrep.2011.03.005

[6] B. Andersson, *Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol.* **7** (1997) 1.

[7] F. James, *Rept. Prog. Phys.* **43**(9) (1980) 1145. http://dx.doi.org/10.1088/0034-4885/43/9/002

[8] Stefan Weinzierl, Introduction to Monte Carlo methods (2000), https://arxiv.org/abs/hep-ph/0006269

[9] K.A. Brueckner, *Phys. Rev.* **86**(1) (1952) 106. http://dx.doi.org/10.1103/PhysRev.86.106

[10] C. Chedester *et al.*, *Phys. Rev.* **82**(6) (1951) 958. http://dx.doi.org/10.1103/PhysRev.82.958

[11] E. Fermi *et al.*, *Phys. Rev.* **85**(5) (1952) 935. http://dx.doi.org/10.1103/PhysRev.85.935

[12] H.L. Anderson *et al.*, *Phys. Rev.* **85**(5) (1952) 936. http://dx.doi.org/10.1103/PhysRev.85.936

[13] H.L. Anderson *et al.*, *Phys. Rev.* **86**(3) (1952) 413. http://dx.doi.org/10.1103/PhysRev.86.413

[14] D.E Nagle, The delta: The first pion nucleon resonance: Its discovery and applications, LALP–84-27, DE84 017107, based on a lecture given at the University of Chicago Symposium in honor of H.L. Anderson, May 11, 1982 (1984).

[15] O.W. Greenberg, *Phys. Rev. Lett.* **13**(20) (1964) 598. http://dx.doi.org/10.1103/PhysRevLett.13.598

[16] M.Y. Han and Y. Nambu, *Phys. Rev.* **139**(4B) (1965) B1006. http://dx.doi.org/10.1103/PhysRev.139.B1006

[17] S.D. Drell and T.-M. Yan, *Phys. Rev. Lett.* **25**(5) (1970) 316. http://dx.doi.org/10.1103/PhysRevLett.25.316

[18] T. Plehn, D. Rainwater and P.Z. Skands, *Phys. Lett.* B **645**(2-3) (2007) 217. http://dx.doi.org/10.1016/j.physletb.2006.12.009

[19] J. Alwall, S. de Visscher and F. Maltoni, *J. High Energy Phys.* **0902** (2009) 017. http://dx.doi.org/10.1088/1126-6708/2009/02/017

[20] A. Papaefstathiou and B. Webber, *J. High Energy Phys.* **06** (2009) 069. http://dx.doi.org/10.1088/1126-6708/2009/06/069

[21] D. Krohn, L. Randall and L.-T. Wang, On the feasibility and utility of ISR tagging (2011), https://arxiv.org/abs/1101.0810

[22] C.K. Vermilion, PhD Thesis, Jet substructure at the Large Hadron Collider: Harder, better, faster, stronger (2011), https://arxiv.org/abs/1101.1335

[23] A. Altheimer *et al.*, *J. Phys. G: Nucl. Part. Phys.* G **39**(6) (2012) 063001. http://dx.doi.org/10.1088/0954-3899/39/6/063001

[24] J. Beringer *et al.*, *Phys. Rev.* D **86**(1) (2012) 010001. http://dx.doi.org/10.1103/PhysRevD.86.010001

[25] S.P. Martin, A Supersymmetry primer (1997), http://dx.doi.org/10.1142/9789812839657_0001

[26] S. Bethke, *Nucl. Phys. Proc. Suppl.* **234** (2013) 229. http://dx.doi.org/10.1016/j.nuclphysbps.2012.12.020

[27] S. Bethke, *Eur. Phys. J.* C **64**(4) (2009) 689. http://dx.doi.org/10.1140/epjc/s10052-009-1173-1

[28] G. Dissertori, A. Gehrmann-De Ridder *et al.*, *J. High Energy Phys.* **0908** (2009) 036. http://dx.doi.org/10.1088/1126-6708/2009/08/036

[29] A. Young, *Proc. London Math. Soc.* **33** (1901) 97.

[30] B.E. Sagan, in *Invariant Theory and Tableaux*, Ed. D. Stanton (Springer, 1990), p. 262. `http://www.mth.msu.edu/users/sagan/Papers/Old/uyt.pdf`.

[31] J. Alwall *et al.*, *J. High Energy Phys.* **1106** (2011) 128. http://dx.doi.org/10.1007/JHEP06(2011)128

[32] A. Pukhov, Calchep 2.3: MSSM, structure functions, event generation, 1, and generation of matrix elements for other packages (2004), https://arxiv.org/abs/hep-ph/0412191

[33] E. Boos *et al.*, *Nucl. Instrum. Meth.* A **534**(1-2) (2004) 50. http://dx.doi.org/10.1016/j.nima.2004.07.096

[34] A. Kanaki and C.G. Papadopoulos, *Comput. Phys. Commun.* **132**(3) (2000) 306. http://dx.doi.org/10.1016/S0010-4655(00)00151-X

[35] F. Krauss, R. Kuhn and G. Soff, *J. High Energy Phys.* **0202** (2002) 044. http://dx.doi.org/10.1088/1126-6708/2002/02/044

[36] M. Moretti, T. Ohl and Jurgen Reuter, O'Mega: An optimizing matrix element generator (2001), https://arxiv.org/abs/hep-ph/0102195

[37] W. Kilian, T. Ohl and J. Reuter, *Eur. Phys. J.* C **71** (2011) 1742. http://dx.doi.org/10.1140/epjc/s10052-011-1742-y

[38] A. Cafarella, C.G. Papadopoulos and M. Worek, *Comput. Phys. Commun.* **180**(10) (2009) 1941. http://dx.doi.org/10.1016/j.cpc.2009.04.023

[39] M. Bähr *et al.*, *Eur. Phys. J.* C **58**(4) (2008) 639. http://dx.doi.org/10.1140/epjc/s10052-008-0798-9

[40] T. Gleisberg and S. Hoeche, *J. High Energy Phys.* **0812** (2008) 039. http://dx.doi.org/10.1088/1126-6708/2008/12/039

[41] T. Sjöstrand. Monte Carlo Generators (2006), https://arxiv.org/abs/hep-ph/0611247

[42] J.C. Collins and D.E. Soper, *Nucl. Phys.* B **194**(3) (1982) 445. http://dx.doi.org/10.1016/0550-3213(82)90021-9

[43] G.F. Sterman. Partons, factorization and resummation, TASI 95 (1995), https://arxiv.org/abs/hep-ph/9606312

[44] R. Brock *et al.*, *Rev. Mod. Phys.* **67**(1) (1995) 157. http://dx.doi.org/10.1103/RevModPhys.67.157

[45] T. Plehn, LHC Phenomenology for Physics Hunters (2008), pp. 125–180, http://dx.doi.org/10.1142/9789812838360_0003

[46] J.C. Collins *et al.*, *Nucl. Phys.* B **250**(1-4) (1985) 199. http://dx.doi.org/10.1016/0550-3213(85)90479-1

[47] G. Altarelli and G. Parisi, *Nucl. Phys.* B **126**(2) (1977) 298. http://dx.doi.org/10.1016/0550-3213(77)90384-4

[48] V.N. Gribov and L.N. Lipatov, *Sov. J. Nucl. Phys.* **15**(4) (1972) 438.

[49] Y.L. Dokshitzer, *Sov. Phys. JETP* **46** (1977) 641.

[50] D. Mason *et al.*, *Phys. Rev. Lett.* **99**(19) (2007) 192001. http://dx.doi.org/10.1103/PhysRevLett.99.192001

[51] A. Cooper-Sarkar, *J. Phys. G: Nucl. Part. Phys.* G **39** (2012) 093001, http://dx.doi.org/10.1088/0954-3899/39/9/093001

[52] S. Alekhin, K. Melnikov and F. Petriello, *Phys. Rev.* D **74**(5) (2006) 054033. http://dx.doi.org/10.1103/PhysRevD.74.054033

[53] E.G. de Oliveira, A.D. Martin and M.G. Ryskin, *Eur. Phys. J.* C **72** (2012) 2069. http://dx.doi.org/10.1140/epjc/s10052-012-2069-z

[54] S. Alekhin *et al.*, The PDF4LHC Working Group Interim Report, (2011), https://arxiv.org/abs/1101.0536

[55] A. Buckley and M. Whalley, HepData reloaded: reinventing the HEP data archive (2010), https://arxiv.org/abs/1006.0517 `http://hepdata.cedar.ac.uk`.

[56] A.D. Martin *et al.*, *Eur. Phys. J.* C **63**(2) (2009) 189. http://dx.doi.org/10.1140/epjc/s10052-009-1072-5

[57] J. Pumplin *et al.*, *J. High Energy Phys.* **0207** (2002) 012. http://dx.doi.org/10.1088/1126-6708/2002/07/012

[58] G. Watt and R.S. Thorne, *J. High Energy Phys.* **1208** (2012) 052. http://dx.doi.org/10.1007/JHEP08(2012)052

[59] M. Bengtsson, T. Sjöstrand and M. van Zijl, *Z. Phys.* C **32**(1) (1986) 67. http://dx.doi.org/10.1007/BF01441353

[60] S. Gieseke, *J. High Energy Phys.* **0501** (2005) 058. http://dx.doi.org/10.1088/1126-6708/2005/01/058

[61] Z. Bern *et al.*, The NLO multileg working group: Summary report, (2008) pp. 83, https://arxiv.org/abs/0803.0494

[62] T. Kinoshita, *J. Math. Phys.* **3**(4) (1962) 650. http://dx.doi.org/10.1063/1.1724268

[63] T.D. Lee and M. Nauenberg, *Phys. Rev.* **133**(6B) (1964) B1549. http://dx.doi.org/10.1103/PhysRev.133.B1549

[64] S. Catani and M.H. Seymour, *Nucl. Phys.* B **485**(1-2) (1997) 291. http://dx.doi.org/10.1016/S0550-3213(96)00589-5

[65] S. Catani and M.H. Seymour, *Phys. Lett.* B **378**(1-4) (1996) 287. http://dx.doi.org/10.1016/0370-2693(96)00425-X

[66] Z. Nagy, *Phys. Rev.* D **68**(9) (2003) 094002. http://dx.doi.org/10.1103/PhysRevD.68.094002

[67] R. Frederix, T. Gehrmann and N. Greiner, *J. High Energy Phys.* **09** (2008) 122. http://dx.doi.org/10.1088/1126-6708/2008/09/122

[68] D.A. Kosower, *Phys. Rev.* D **57**(9) (1998) 5410. http://dx.doi.org/10.1103/PhysRevD.57.5410

[69] D.A. Kosower. *Phys. Rev.* D **71**(4) (2005) 045016. http://dx.doi.org/10.1103/PhysRevD.71.045016

[70] A. Gehrmann-De Ridder, T. Gehrmann and E.W.N. Glover, *J. High Energy Phys.* **09** (2005) 056. http://dx.doi.org/10.1088/1126-6708/2005/09/056

[71] A. Gehrmann-De Ridder *et al.*, *J. High Energy Phys.* **0712** (2007) 094. http://dx.doi.org/10.1088/1126-6708/2007/12/094

[72] S. Weinzierl, *Phys. Rev. Lett.* **101**(16) (2008) 162001.

http://dx.doi.org/10.1103/PhysRevLett.101.162001

[73] G. Heinrich. *Int. J. Mod. Phys.* A **23**(10) (2008) 1457. http://dx.doi.org/10.1142/S0217751X08040263

[74] R. Boughezal, K. Melnikov and F. Petriello, *Phys. Rev.* D **85**(3) (2012) 034025. http://dx.doi.org/10.1103/PhysRevD.85.034025

[75] S. Catani and M. Grazzini, *Phys. Rev. Lett.* **98**(22) (2007) 222002. http://dx.doi.org/10.1103/PhysRevLett.98.222002

[76] G. Gustafson and U. Pettersson, *Nucl. Phys.* B **306**(4) (1988) 746. http://dx.doi.org/10.1016/0550-3213(88)90441-5

[77] A. Gehrmann-De Ridder and M. Ritzmann, *J. High Energy Phys.* **0907** (2009) 041. http://dx.doi.org/10.1088/1126-6708/2009/07/041

[78] A. Gehrmann-De Ridder, M. Ritzmann and P. Skands, *Phys. Rev.* D **85**(1) (2012) 014013. http://dx.doi.org/10.1103/PhysRevD.85.014013

[79] A.J. Larkoski and M.E. Peskin, *Phys. Rev.* D **81**(5) (2010) 054010. http://dx.doi.org/10.1103/PhysRevD.81.054010

[80] W.T. Giele, D.A. Kosower and P.Z. Skands, *Phys. Rev.* D **84**(5) (2011) 054003. http://dx.doi.org/10.1103/PhysRevD.84.054003

[81] J.J. Lopez-Villarejo and P. Skands, *J. High Energy Phys.* **1111** (2011) 150. http://dx.doi.org/10.1007/JHEP11(2011)150

[82] A. Banfi, G.P. Salam and G. Zanderighi, *J. High Energy Phys.* **1006** (2010) 038. http://dx.doi.org/10.1007/JHEP06(2010)038

[83] V. Khachatryan *et al.*, *Phys. Lett.* B **699**(1-2) (2011) 48. http://dx.doi.org/10.1016/j.physletb.2011.03.060

[84] G. Aad *et al.*, *Eur. Phys. J.* C **72** (2012) 2211. http://dx.doi.org/10.1140/epjc/s10052-012-2211-y

[85] K. Wraight and P. Skands, *Eur. Phys. J.* C **71** (2011) 1628. http://dx.doi.org/10.1140/epjc/s10052-011-1628-z

[86] M. Cacciari, G.P. Salam and G. Soyez. *J. High Energy Phys.* **0804** (2008) 063. http://dx.doi.org/10.1088/1126-6708/2008/04/063

[87] A. Abdesselam *et al.*, *Eur. Phys. J.* C **71** (2011) 1661. http://dx.doi.org/10.1140/epjc/s10052-011-1661-y

[88] M. Cacciari, G.P. Salam and S. Sapeta, *J. High Energy Phys.* **1004** (2010) 065. http://dx.doi.org/10.1007/JHEP04(2010)065

[89] G. Corcella *et al.*, *J. High Energy Phys.* **01** (2001) 010. http://dx.doi.org/10.1088/1126-6708/2001/01/010

[90] T. Sjöstrand, S. Mrenna and P. Skands, *J. High Energy Phys.* **05** (2006) 026. http://dx.doi.org/10.1088/1126-6708/2006/05/026

[91] T. Sjöstrand *et al.*, An Introduction to PYTHIA 8.2 (2014), http://doi.org/10.1016/j.cpc.2015.01.024

[92] T. Gleisberg *et al.*, *J. High Energy Phys.* **02** (2009) 007. http://dx.doi.org/10.1088/1126-6708/2009/02/007

[93] E. Boos *et al.*, Generic user process interface for event generators (2001), https://arxiv.org/abs/hep-ph/0109068

[94] J. Alwall *et al.*, *Comput. Phys. Commun.* **176**(4) (2007) 300. http://dx.doi.org/10.1016/j.cpc.2006.11.010

[95] A. Buckley *et al.*, Rivet user manual (2010), https://arxiv.org/abs/1003.0694

[96] S. Agostinelli *et al.*, *Nucl. Instrum. Meth.* A **506**(3) (2003) 250.

http://dx.doi.org/10.1016/S0168-9002(03)01368-8

[97] M. Dobbs and J.B. Hansen, *Comput. Phys. Commun.* **134**(1) (2001) 41. http://dx.doi.org/10.1016/S0010-4655(00)00189-2

[98] T. Sjöstrand, Monte Carlo Tools (2009), https://arxiv.org/abs/0911.5286

[99] R. Kleiss, W.J. Stirling and S.D. Ellis, *Comput. Phys. Commun.* **40**(2-3) (1986) 359. http://dx.doi.org/10.1016/0010-4655(86)90119-0

[100] G.P. Lepage, *J. Comput. Phys.* **27**(2) (1978) 192. http://dx.doi.org/10.1016/0021-9991(78)90004-9

[101] G.P. Lepage, VEGAS – An adaptive multi-dimensional integration program, CLNS-80/447 (1980), http://inspirehep.net/record/153221/

[102] P.D. Draggiotis, A. van Hameren and R. Kleiss, *Phys. Lett.* B **483**(1-3) (2000) 124. http://dx.doi.org/10.1016/S0370-2693(00)00532-3

[103] G. Marchesini and B.R. Webber, *Nucl. Phys.* B **238**(1) (1984) 1. http://dx.doi.org/10.1016/0550-3213(84)90463-2

[104] G. Marchesini and B.R. Webber, *Nucl. Phys.* B **310**(3-4) (1988) 461. http://dx.doi.org/10.1016/0550-3213(88)90089-2

[105] S. Gieseke, P. Stephens and B. Webber, *J. High Energy Phys.* **12** (2003) 045. http://dx.doi.org/10.1088/1126-6708/2003/12/045

[106] M. Bengtsson and T. Sjöstrand, *Nucl. Phys.* B **289** (1987) 810. http://dx.doi.org/10.1016/0550-3213(87)90407-X

[107] L. Lönnblad, *Comput. Phys. Commun.* **71**(1-2) (1992) 15. http://dx.doi.org/10.1016/0010-4655(92)90068-A

[108] Z. Nagy and D.E. Soper, *J. High Energy Phys.* **10** (2005) 024. http://dx.doi.org/10.1088/1126-6708/2005/10/024

[109] S. Schumann and F. Krauss, *J. High Energy Phys.* **0803** (2008) 038. http://dx.doi.org/10.1088/1126-6708/2008/03/038

[110] W.T. Giele, D.A. Kosower and P.Z. Skands, *Phys. Rev.* D **78**(1) (2008) 014026. http://dx.doi.org/10.1103/PhysRevD.78.014026

[111] T. Sjöstrand and P.Z. Skands, *Eur. Phys. J.* C **39**(2) (2005) 129. http://dx.doi.org/10.1140/epjc/s2004-02084-y

[112] N. Fischer *et al.*, *Eur. Phys. J.* C **74** (2014) 2831. http://dx.doi.org/10.1140/epjc/s10052-014-2831-5

[113] J.C. Collins, *Nucl. Phys.* B **304** (1988) 794. http://dx.doi.org/10.1016/0550-3213(88)90654-2

[114] I.G. Knowles, *Nucl. Phys.* B **310**(3-4) (1988) 571. http://dx.doi.org/10.1016/0550-3213(88)90092-2

[115] P. Richardson, *J. High Energy Phys.* **11** (2001) 029. http://dx.doi.org/10.1088/1126-6708/2001/11/029

[116] S. Catani, B.R. Webber and G. Marchesini, *Nucl. Phys.* B **349**(3) (1991) 635. http://dx.doi.org/10.1016/0550-3213(91)90390-J

[117] M.A. Gigg and P. Richardson, Simulation of finite width effects in physics beyond the standard model (2008), https://arxiv.org/abs/0805.3037

[118] M.H. Seymour, *Phys. Lett.* B **354**(3-4) (1995) 409. http://dx.doi.org/10.1016/0370-2693(95)00699-L

[119] T. Stelzer and S. Willenbrock, *Phys. Lett.* B **374**(1-3) (1996) 169. http://dx.doi.org/10.1016/0370-2693(96)00178-5

[120] S.J. Parke and Y. Shadmi, *Phys. Lett.* B **387**(1) (1996) 199.

http://dx.doi.org/10.1016/0370-2693(96)00998-7

[121] J.M. Smillie and B.R. Webber, *J. High Energy Phys.* **0510** (2005) 069. http://dx.doi.org/10.1088/1126-6708/2005/10/069

[122] M.H. Seymour, *Nucl. Phys.* B **436**(3) (1995) 443. http://dx.doi.org/10.1016/0550-3213(94)00554-R

[123] M.H. Seymour, *Comp. Phys. Commun.* **90**(1) (1995) 95. http://dx.doi.org/10.1016/0010-4655(95)00064-M

[124] S. Catani *et al.*, *J. High Energy Phys.* **11** (2001) 063. http://dx.doi.org/10.1088/1126-6708/2001/11/063

[125] L. Lönnblad, *J. High Energy Phys.* **05** (2002) 046. http://dx.doi.org/10.1088/1126-6708/2002/05/046

[126] N. Lavesson and L. Lönnblad, *J. High Energy Phys.* **07** (2005) 054. http://dx.doi.org/10.1088/1126-6708/2005/07/054

[127] M.L. Mangano *et al.*, *J. High Energy Phys.* **01** (2007) 013. http://dx.doi.org/10.1088/1126-6708/2007/01/013

[128] S. Mrenna and P. Richardson, *J. High Energy Phys.* **05** (2004) 040. http://dx.doi.org/10.1088/1126-6708/2004/05/040

[129] B. Cooper *et al.*, *Eur. Phys. J.* C **72** (2012) 2078. http://dx.doi.org/10.1140/epjc/s10052-012-2078-y

[130] N. Lavesson and L. Lönnblad, *J. High Energy Phys.* **12** (2008) 070. http://dx.doi.org/10.1088/1126-6708/2008/12/070

[131] S. Frixione and B.R. Webber, *J. High Energy Phys.* **06** (2002) 029. http://dx.doi.org/10.1088/1126-6708/2002/06/029

[132] S. Frixione, P. Nason, and B.R. Webber, *J. High Energy Phys.* **08** (2003) 007. http://dx.doi.org/10.1088/1126-6708/2003/08/007

[133] S. Frixione and B.R. Webber, The MC@NLO 3.4 event generator (2008), https://arxiv.org/abs/0812.0770

[134] S. Frixione, P. Nason and C. Oleari, *J. High Energy Phys.* **11** (2007) 070. http://dx.doi.org/10.1088/1126-6708/2007/11/070

[135] K. Hamilton and P. Nason, *J. High Energy Phys.* **06** (2010) 039. http://dx.doi.org/10.1007/JHEP06(2010)039

[136] M. Bengtsson and T. Sjöstrand, *Phys. Lett.* B **185**(3-4) (1987) 435. http://dx.doi.org/10.1016/0370-2693(87)91031-8

[137] S. Alioli *et al.*, *J. High Energy Phys.* **06** (2010) 043. http://dx.doi.org/10.1007/JHEP06(2010)043

[138] L. Hartgring, E. Laenen and P. Skands, *J. High Energy Phys.* **1310** (2013) 127. http://dx.doi.org/10.1007/JHEP10(2013)127

[139] P.Z. Skands, QCD for Collider Physics, 18th European School on High-Energy Physics (ESHEP), Raseborg, Finland, 2010, arXiv:1202.1629 (2011).

[140] G.S. Bali and K. Schilling, *Phys. Rev.* D **46**(6) (1992) 2636. http://dx.doi.org/10.1103/PhysRevD.46.2636

[141] D. Amati and G. Veneziano, *Phys. Lett.* B **83**(1) (1979) 87. http://dx.doi.org/10.1016/0370-2693(79)90896-7

[142] S. Navin, Diffraction in Pythia (2010), https://arxiv.org/abs/1005.3894

[143] X. Artru and G. Mennessier, *Nucl. Phys.* B **70**(1) (1974) 93. http://dx.doi.org/10.1016/0550-3213(74)90360-5

[144] M. Travis *et al.*, *Sixteen Tons*, originally recorded for Folk Songs of the Hills (Capitol Records,

1946).

[145] B. Andersson, G. Gustafson and T. Sjöstrand, *Nucl. Phys.* B **197**(1) (1982) 45. http://dx.doi.org/10.1016/0550-3213(82)90153-5

[146] B. Andersson, G. Gustafson and T. Sjöstrand, *Phys. Scripta* **32**(6) (1985) 574. http://dx.doi.org/10.1088/0031-8949/32/6/003

[147] P. Eden and G. Gustafson, *Z. Phys.* C **75**(1) (1997) 41. http://dx.doi.org/10.1007/s002880050445

[148] T. Sjöstrand and P.Z. Skands, *Nucl. Phys.* B **659**(1-2) (2003) 243. http://dx.doi.org/10.1016/S0550-3213(03)00193-7

[149] T. Sjöstrand and P.Z. Skands, *J. High Energy Phys.* **03** (2004) 053. http://dx.doi.org/10.1088/1126-6708/2004/03/053

[150] B. Andersson, G. Gustafson and B. Söderberg, *Z. Phys.* C **20**(4) (1983) 317. http://dx.doi.org/10.1007/BF01407824

[151] M.G. Bowler, *Z. Phys.* C **11**(2) (1981) 169. http://dx.doi.org/10.1007/BF01574001

[152] K.G. Wilson, *Phys. Rev.* D **10**(8) (1974) 2445. http://dx.doi.org/10.1103/PhysRevD.10.2445

[153] G. Antchev *et al.*, *Phys. Rev. Lett.* **111**(1) (2013) 012001. http://dx.doi.org/10.1103/PhysRevLett.111.012001

[154] A. Donnachie and P.V. Landshoff, *Phys. Lett.* B **727**(4-5) (2013) 500. http://dx.doi.org/10.1016/j.physletb.2013.10.068

[155] V.A. Khoze *et al.*, *Eur. Phys. J.* C **69**(1) (2010) 85. http://dx.doi.org/10.1140/epjc/s10052-010-1392-5

[156] B. Blok *et al.*, *Phys. Rev.* D **83**(7) (2011) 071501. http://dx.doi.org/10.1103/PhysRevD.83.071501

[157] B. Blok *et al.*, *Eur. Phys. J.* C **72** (2012) 1963. http://dx.doi.org/10.1140/epjc/s10052-012-1963-8

[158] J.R. Gaunt and W.J. Stirling, Single and double perturbative splitting diagrams in double parton scattering (2012), https://arxiv.org/abs/1202.3056

[159] A.V. Manohar and W.J. Waalewijn, *Phys. Rev.* D **85**(11) (2012) 114009. http://dx.doi.org/10.1103/PhysRevD.85.114009

[160] T. Sjöstrand and M. van Zijl, *Phys. Rev.* D **36**(7) (1987) 2019. http://dx.doi.org/10.1103/PhysRevD.36.2019

[161] P. Skands, S. Carrazza and J. Rojo, *Eur. Phys. J.* C **74(8)** (2014) 3024. http://dx.doi.org/10.1140/epjc/s10052-014-3024-y

[162] J.M. Butterworth, J.R. Forshaw and M.H. Seymour, *Z. Phys.* C **72**(4) (1996) 637. http://dx.doi.org/10.1007/s002880050286

[163] M. Bähr *et al.*, Soft interactions in Herwig++ (2009), https://arxiv.org/abs/0905.4671

[164] M.G. Ryskin, A.D. Martin and V.A. Khoze, *Eur. Phys. J.* C **71** (2011) 1617. http://dx.doi.org/10.1140/epjc/s10052-011-1617-2

[165] H. Schulz and P.Z. Skands, *Eur. Phys. J.* C **71** (2011) 1644. http://dx.doi.org/10.1140/epjc/s10052-011-1644-z

[166] P.Z. Skands, *Phys. Rev.* D **82**(7) (2010) 074018. http://dx.doi.org/10.1103/PhysRevD.82.074018

[167] P.Z. Skands, Soft-QCD and UE spectra in pp collisions at very high CM energies (a Snowmass white paper) (2013), https://arxiv.org/abs/1308.2813

[168] R. Corke and T. Sjöstrand, *J. High Energy Phys.* **1105** (2011) 009. http://dx.doi.org/10.1007/JHEP05(2011)009

[169] G. Aad *et al.*, *Phys. Rev.* D **83**(11) (2011) 112001, http://dx.doi.org/10.1103/PhysRevD.83.112001

[170] A. Karneyeu *et al.*, *Eur. Phys. J.* C **74** (2014) 2714.

http://dx.doi.org/10.1140/epjc/s10052-014-2714-9

[171] D. Lombrana Gonzalez *et al.*, *Conf. Proc.* C **1205201** (2012) 505, https://cds.cern.ch/record/1463291

[172] A. Buckley *et al.*, *Eur. Phys. J.* C **65** (2010) 331. http://dx.doi.org/10.1140/epjc/s10052-009-1196-7

[173] P. Richardson and D. Winn, *Eur. Phys. J.* C **72** (2012) 2178. http://dx.doi.org/10.1140/epjc/s10052-012-2178-8

[174] M. Ritzmann, D.A. Kosower and P. Skands, *Phys. Lett.* B **718**(4-5) (2013) 1345. http://dx.doi.org/10.1016/j.physletb.2012.12.003

[175] P. Skands, B. Webber and J. Winter, *J. High Energy Phys.* **1207** (2012) 151. http://dx.doi.org/10.1007/JHEP07(2012)151

[176] R. Field, *Acta Phys. Polon.* B **42** (2011) 2631. http://dx.doi.org/10.5506/APhysPolB.42.2631

[177] L. Lönnblad and T. Sjöstrand, *Eur. Phys. J.* C **2** (1998) 165. http://dx.doi.org/10.1007/s100520050131

[178] R. Corke and T. Sjöstrand, *J. High Energy Phys.* **01** (2009) 035. http://dx.doi.org/10.1007/JHEP01(2010)035

[179] J. Rathsman, *Phys. Lett.* B **452**(3-4) (1999) 364. http://dx.doi.org/10.1016/S0370-2693(99)00291-9

[180] P.Z. Skands and D. Wicke, *Eur. Phys. J.* C **52**(1) (2007) 133. http://dx.doi.org/10.1140/epjc/s10052-007-0352-1

[181] S. Gieseke, C. Rohr and A. Siodmok, *Eur. Phys. J.* C **72** (2012) 2225. http://dx.doi.org/10.1140/epjc/s10052-012-2225-5

[182] K. Werner, I. Karpenko and T. Pierog, *Acta Phys. Polon. Supp.* **4** (2011) 629. http://dx.doi.org/10.5506/APhysPolBSupp.4.629

[183] N. Ellis *et al.*, Proc. 2008 European School of High-Energy Physics (ESHEP), Herbeumont-sur-Semois, Belgium, (2009), CERN-2009-002 (2009). http://dx.doi.org/10.5170/CERN-2009-002

[184] C. Grojean *et al.*, Proc. 2009 European School of High-Energy Physics (ESHEP), Bautzen, Germany, 2009 (2010).

[185] N. Ellis *et al.*, 2007 European School of High-Energy Physics, Trest, Czech Republic, 2007, CERN-2008-007 (2008). http://dx.doi.org/10.5170/CERN-2008-007

Proceedings of the 2014 Asia–Europe–Pacific School of High-Energy Physics, Puri, India, 4–17 November 2014, edited by
M. Mulders and R. Godbole, CERN Yellow Reports: School Proceedings, Vol. 2/2017, CERN-2017-005-SP (CERN, Geneva, 2017)

Flavour Physics and CP Violation

S. J. Lee and H. Serôdio
Department of Physics, Korea Advanced Institute of Science and Technology, Daejeon, Korea

Abstract
We present the invited lectures given at the second Asia-Europe-Pacific School of High-Energy Physics (AEPSHEP), which took place in Puri, India in November 2014. The series of lectures aimed at graduate students in particle experiment/theory, covering the the very basics of flavour physics and CP violation, some useful theoretical methods such as OPE and effective field theories, and some selected topics of flavour physics in the era of LHC.

Keywords
Lectures; flavor; CP violation; CKM matrix; flavor changing neutral currents; GIM mechanism.

1 Short introduction

We present the invited lectures given at the second Asia-Europe-Pacific School of High-Energy Physics (AEPSHEP), which took place in Puri, India in November 2014. The physics background of students attending the school are diverse as some of them were doing their PhD studies in experimental particle physics, others in theoretical particle physics. The lectures were planned and organized, such that students from different background can still get benefit from basic topics of broad interest in a modern way, trying to explain otherwise complicated concepts necessary to know for understanding the current ongoing researches in the field, in a relatively simple language from first principles.

These notes present a small compilation of several results that over the years has become standard in particle physics, and more concretely in the area of flavour physics. These are by no means a complete and self-contained course in flavour physics, but rather a brief introduction to several topics that should be explored in more detail by additional references for the interested readers. For the topics addressed in these notes there are several textbooks and review articles that have become standard references; here we compile an incomplete list:

- For aspects concerning the building blocks of gauge theories and the standard model see, for example, [1]
- For CP and flavour aspects in particle physics the books [2–4] are two excellent sources, as well as the more specific reviews [5–18]
- For topics related with effective field theories we refer the reader to [19–23]

2 The building blocks in particle physics

2.1 What is flavour and why do we care?

In Particle Physics one attributes quantum numbers to particles in order to classify them as representations of the symmetries describing the dynamics of the underlying model. This classification allows us to extract a lot of information just from first principles. In nature there are several copies of the same fermionic gauge representation, i.e. several fields that are assigned the same quantum numbers. We then say that different copies belong to different flavours (or families). Flavour physics describes the interactions that distinguish between flavours, i.e. between the different copies.

The fermions can interact through pure gauge interactions. These interaction are related to the unbroken symmetries and mediated therefore by massless gauge bosons. They do not distinguish among

https://doi.org/10.23730/CYRSP-2017-002.125

the flavours and do not constitute part of flavour physics. Fermions can also have Yukawa interactions, i.e. interactions where two fermions couple to a scalar. These interactions are source of flavour and CP violation. Within the Standard Model (SM), flavour physics refers to the weak and Yukawa interactions.

Flavour physics can predict new physics (NP) before it's directly observed. Some examples are:

- The smallness of $\Gamma(K_L \to \mu^+\mu^-)/\Gamma(K^+ \to \mu^+\nu)$ allowed for the prediction of the charm quark
- The size of Δm_K allowed for the charm mass prediction
- The measurement of ϵ_K allowed for the prediction of the third generation
- The size of Δm_B allowed for a quite accurate top mass prediction ($\sim 150\,\text{GeV}$)
- The measurement of neutrino flavour transitions led to the discovery of neutrino masses

2.2 Discrete symmetries in particle physics

In this section we present the discrete symmetries C, P and T, which play a leading role in the construction of the present model of particle physics. These three symmetries do not leave, separately, the SM Lagrangian invariant but their product CPT does (at least everything points on that direction). These discrete symmetries give rise to multiplicative conservation laws. They have three levels of action: on the particle states, on the creation and annihilation operators, and on the fields. The action on one level determines the action on the other two. The main properties of these symmetries are:

- **Charge Conjugation**
  Charge conjugation on the states reverses the quantum numbers of particles that are associated with internal symmetries. The charge conjugate of a particle is another particle with the same energy and momentum but opposite charges (anti-particle). Charge conjugation on the fields converts a field $\psi(x)$ into a field $\psi^c(x)$ with opposite internal quantum numbers. If charge conjugation is a symmetry of the quantum field theory, there must exist a unitary operator $\mathcal{C}$ which represents it.
  We can use charge conjugation in order to eliminate final states for scattering and decay processes and to provide a link between different processes involving charged particles.
- **Parity**
  Classical parity is any element in the component of the Lorentz group that contains the matrix $P = \text{diag}(1, -1, -1, -1)$. Parity, like charge conjugation, gives rise to a multiplicative conservation law. For example, the η meson and the pions are pseudoscalars (eigenstates with eigenvalue -1 as opposed "$+1$ for scars), and so the decay $\eta \to \pi^+\pi^-$ is forbidden by conservation of parity. However, since parity transforms space, the eigenvalues of parity depend on the orbital angular momentum of a state and the intrinsic parity of a state is not in general conserved.
- **Time Reversal**
  The idea of time reversal is to take the time evolution of some system and reverse it. To separate the effects of charge conjugation from those of time reversal, it is customary to assume that time reversal preserves the internal quantum numbers of all particles. In classical mechanics, time reversal can be implemented by changing the sign of the Hamiltonian. If we suppose that this effect is achieved in quantum theory by a unitary transformation U_T, we get

  $$U_T^\dagger e^{iHt} U_T = e^{iHt} \quad \Rightarrow \quad U_T^\dagger H U_T = -H \quad \Rightarrow HU_T|n\rangle = -E_n U_T|n\rangle\,, \tag{1}$$

  for any state $|n\rangle$, entering in conflict with the principle that energy should be bounded from below. The way to solve this is by dropping the unitary operator and represent time reversal by an anti-unitary operator operator $\mathcal{T}$.

Tables 1–2 summarize some of the most important transformations under these symmetries.

Table 1: Discrete symmetry transformations for photon, gluon, complex scalar and fermion fields. We have defined: $\psi^c = C\overline{\psi}^T$ and s^a is $+1$ for $a = 1, 3, 4, 6, 8$ while -1 for $a, 2, 5, 7$.

Fields transformations		
Photon:	Gluon:	Complex scalar:
$\mathcal{P}A_\mu(t,\vec{r})\mathcal{P}^\dagger = A^\mu(t,-\vec{r})$	$\mathcal{P}G^a_\mu(t,\vec{r})\mathcal{P}^\dagger = G^{a\mu}(t,-\vec{r})$	$\mathcal{P}\phi(t,\vec{r})\mathcal{P}^\dagger = e^{i\alpha_p}\phi(t,-\vec{r})$
$\mathcal{T}A_\mu(t,\vec{r})\mathcal{T}^{-1} = A^\mu(-t,\vec{r})$	$\mathcal{T}G^a_\mu(t,\vec{r})\mathcal{T}^{-1} = s^a G^{a\mu}(-t,\vec{r})$	$\mathcal{T}\phi(t,\vec{r})\mathcal{T}^{-1} = e^{i\alpha_t}\phi(-t,\vec{r})$
$\mathcal{C}A_\mu(t,\vec{r})\mathcal{C}^\dagger = -A_\mu(t,\vec{r})$	$\mathcal{C}G^a_\mu(t,\vec{r})\mathcal{C}^\dagger = -sG^a_\mu(t,\vec{r})$	$\mathcal{C}\phi(t,\vec{r})\mathcal{C}^\dagger = e^{i\alpha_c}\phi^\dagger(t,\vec{r})$
$\mathcal{CP}A_\mu(t,\vec{r})\mathcal{CP}^\dagger = -A_\mu(t,-\vec{r})$	$\mathcal{CP}G^a_\mu(t,\vec{r})\mathcal{CP}^\dagger = -s^aG^a_\mu(t,-\vec{r})$	$\mathcal{CP}\phi(t,\vec{r})\mathcal{CP}^\dagger = e^{i\alpha}\phi^\dagger(t,-\vec{r})$

Fermion:

$\mathcal{P}\psi(t,\vec{r})\mathcal{P}^\dagger = e^{i\beta_p}\gamma^0\psi(t,-\vec{r})$	$\mathcal{P}\overline{\psi}(t,\vec{r})\mathcal{P}^\dagger = e^{-i\beta_p}\overline{\psi}(t,-\vec{r})\gamma^0$
$\mathcal{T}\psi(t,\vec{r})\mathcal{T}^{-1} = e^{i\beta_t}\gamma^*_0\gamma^*_5 C^*\overline{\psi}^\dagger(-t,\vec{r})$	$\mathcal{T}\overline{\psi}(t,\vec{r})\mathcal{T}^{-1} = e^{-i\beta_t}\psi^\dagger(-t,\vec{r})(C^{-1})^*\gamma^*_5\gamma^*_0$
$\mathcal{C}\psi(t,\vec{r})\mathcal{C}^\dagger = e^{i\beta_c}\phi^c(t,\vec{r})$	$\mathcal{C}\overline{\psi}(t,\vec{r})\mathcal{C}^\dagger = e^{i\beta_c}\overline{\psi^c}(t,\vec{r})$
$\mathcal{CP}\psi(t,\vec{r})\mathcal{CP}^\dagger = e^{i\alpha}\gamma^0 C\overline{\psi}^T(t,-\vec{r})$	$\mathcal{CP}\overline{\psi}(t,\vec{r})\mathcal{CP}^\dagger = e^{-i\alpha}\psi^T(t,-\vec{r})C^{-1}\gamma^0$

Table 2: Symmetry transformation properties of some fermionic bilinears under the action of discrete symmetries. Overall phases and the coordinates have been omitted.

Bilinear	$\mathcal{P}$	$\mathcal{T}$	$\mathcal{C}$	$\mathcal{CP}$	$\mathcal{CPT}$
$\overline{\psi}\chi$	$\overline{\psi}\chi$	$\overline{\psi}\chi$	$\overline{\chi}\psi$	$\overline{\chi}\psi$	$\overline{\chi}\psi$
$\overline{\psi}\gamma_5\chi$	$-\overline{\psi}\gamma_5\chi$	$\overline{\psi}\gamma_5\chi$	$\overline{\chi}\gamma_5\psi$	$-\overline{\chi}\gamma_5\psi$	$-\overline{\chi}\gamma_5\psi$
$\overline{\psi}P_{L,R}\chi$	$\overline{\psi}P_{R,L}\chi$	$\overline{\psi}P_{L,R}\chi$	$\overline{\chi}P_{L,R}\psi$	$\overline{\chi}P_{R,L}\psi$	$\overline{\chi}P_{R,L}\psi$
$\overline{\psi}\gamma^\mu\chi$	$\overline{\psi}\gamma_\mu\chi$	$\overline{\psi}\gamma_\mu\chi$	$-\overline{\chi}\gamma^\mu\psi$	$-\overline{\chi}\gamma_\mu\psi$	$-\overline{\chi}\gamma^\mu\psi$
$\overline{\psi}\gamma^\mu\gamma_5\chi$	$-\overline{\psi}\gamma_\mu\gamma_5\chi$	$\overline{\psi}\gamma_\mu\gamma_5\chi$	$\overline{\chi}\gamma^\mu\gamma_5\psi$	$-\overline{\chi}\gamma_\mu\gamma_5\psi$	$-\overline{\chi}\gamma^\mu\gamma_5\psi$
$\overline{\psi}\gamma^\mu P_{L,R}\chi$	$\overline{\psi}\gamma_\mu P_{R,L}\chi$	$\overline{\psi}\gamma_\mu P_{L,R}\chi$	$-\overline{\chi}\gamma^\mu P_{R,L}\psi$	$-\overline{\chi}\gamma_\mu P_{L,R}\psi$	$-\overline{\chi}\gamma^\mu P_{L,R}\psi$
$\overline{\psi}\sigma^{\mu\nu}\chi$	$\overline{\psi}\sigma_{\mu\nu}\chi$	$-\overline{\psi}\sigma_{\mu\nu}\chi$	$-\overline{\chi}\sigma^{\mu\nu}\psi$	$-\overline{\chi}\sigma_{\mu\nu}\psi$	$\overline{\chi}\sigma^{\mu\nu}\psi$

2.3 Basic Building Blocks of the SM

In this section we shall briefly present the building blocks of the SM, taking special attention to the relevant sector for flavour physics. Modern Quantum Field Theories are based on the gauge principle: *The Lagrangian is invariant under a continuous group of local transformations. For each group generator there necessarily arises a corresponding vector field called the gauge field, responsible for ensuring the Lagrangian invariance under the local group transformations.*

Following the above principle, modern theories are developed through three simple steps:

(1) Define the gauge symmetry

(2) Choose the representations of the matter content under the symmetry

(3) Choose the way your original symmetry is broken

The first two steps define the model in the unbroken phase. We then need a way to break this symmetry since at low energies we know that only charge (and colour) is manifestly preserved.

The best example satisfying the above three conditions and having an enormous success when confronting with data is the SM. The model construct upon the gauge group (step **(1)**)

$$\mathcal{G}_{\rm SM} = SU(3)_C \times SU(2)_L \times U(1)_Y \,. \tag{2}$$

From the gauge principle, each generator of $\mathcal{G}_{\rm SM}$ has a associated gauge vector field (first four lines of the table on the right in Table 3). The known matter fields are embedded in irreducible representations of $\mathcal{G}_{\rm SM}$ (step **(2)**) and are presented on the left table in Table 3. The gauge fields interact with matter

Table 3: Standard model particle content, symmetry representations and forces.

Matter	**Flavour**	$\mathcal{G}_{\rm SM}$
$q_{L\alpha} \equiv \begin{pmatrix} u_{L\alpha} \\ d_{L\alpha} \end{pmatrix}$	$\begin{pmatrix} \mathrm{u}_L \\ \mathrm{d}_L \end{pmatrix}$, $\begin{pmatrix} \mathrm{c}_L \\ \mathrm{s}_L \end{pmatrix}$, $\begin{pmatrix} \mathrm{t}_L \\ \mathrm{b}_L \end{pmatrix}$	$(\mathbf{3}, \mathbf{2}, 1/6)$
$u_{R\alpha}$	u_R , c_R , t_R	$(\mathbf{3}, \mathbf{1}, 2/3)$
$d_{R\alpha}$	d_R , s_R , b_R	$(\mathbf{3}, \mathbf{1}, -1/3)$
$\ell_{L\alpha} \equiv \begin{pmatrix} \nu_{L\alpha} \\ e_{L\alpha} \end{pmatrix}$	$\begin{pmatrix} \nu_{L\mathrm{e}} \\ \mathrm{e}_L \end{pmatrix}$, $\begin{pmatrix} \nu_{L\mu} \\ \mu_L \end{pmatrix}$, $\begin{pmatrix} \nu_{L\tau} \\ \tau_L \end{pmatrix}$	$(\mathbf{1}, \mathbf{2}, -1/2)$
$e_{R\alpha}$	e_R , μ_R , τ_R	$(\mathbf{1}, \mathbf{1}, -1)$

Bosons	**Force**
G^a_μ	Strong
$W^\pm_\mu$, Z^0_μ	Weak
A_μ	EM
$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}$	Yukawa-type $(\mathbf{1}, \mathbf{2}, 1/2)$

through the covariant derivative, which can be expressed in terms of the physical gauge bosons as

$$D_\mu = \partial_\mu - i g_s G^a_\mu \frac{\lambda_a}{2} - ig\left(W^+_\mu T_+ + W^-_\mu T_-\right) - ieA_\mu Q - \frac{ig}{c_W} Z^0_\mu \left(T_3 - s^2_W Q\right) , \tag{3}$$

with $(T_\pm)_{ij} = (|\epsilon_{ij}| \pm \epsilon_{ij})/(2\sqrt{2})$ and $(T_3)_{ij} = \delta_{ij}(-1)^{ij}/2$ for the $SU(2)$ doublet representations. The electric charge Q is a linear combination of the generator of $U(1)_Y$ and the diagonal generator of $SU(2)_L$, and reads $Q = Y + T_3$. The full SM Lagrangian is now a combination of several "distinct" parts which can, in many scenarios, be studied separately. We write it as

$$\mathcal{L}_{\rm SM} = \mathcal{L}^{\rm gauge}_{\rm Kin} + \mathcal{L}^{\rm fermion}_{\rm Kin} + \mathcal{L}_{\rm Higgs} + \mathcal{L}_{\rm Yukawa} + \mathcal{L}_{\rm gf} + \mathcal{L}_{\rm FP} \, . \tag{4}$$

The terms $\mathcal{L}_{\rm gf}$ and $\mathcal{L}_{\rm FP}$ denote the gauge fixing and Faddeev-Popov Lagrangian, respectively. While these contributions are very important for the self-consistency of the model, for flavour physics they play no role and, therefore, shall be ignored in these notes. The other Lagrangian terms are presented in Table 4. A useful summary of Feynman rules for the SM can be found in [24].

Table 4: Standard model Lagrangian equations for the four relevant sectors. With the following definitions: $G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g_s f^{abc} G^a_\mu G^b_\nu \, , (a,b,c = 1,...,8)$, $W^a_{\mu\nu} = \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g\epsilon^{abc} W^a_\mu W^b_\nu \, (a,b,c = 1,...,3)$, $B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$, $Y^{u,d,\ell}$ the up, down and charged-lepton Yuwaka coupling matrices and $\tilde{\phi} = i\tau_2 \phi^*$.

Sector	**Lagrangian**
$\mathcal{L}^{\rm gauge}_{\rm kin}$	$-\frac{1}{4} G^{a\mu\nu} G^a_{\mu\nu} - \frac{1}{4} W^{a\mu\nu} W^a_{\mu\nu} - \frac{1}{4} B^{\mu\nu} B_{\mu\nu}$
$\mathcal{L}^{\rm fermion}_{\rm kin}$	$\overline{q^0_{L\alpha}} i \not{D} q^0_{L\alpha} + \overline{u^0_{R\alpha}} i \not{D} u^0_{R\alpha} + \overline{d^0_{R\alpha}} i \not{D} d^0_{R\alpha} + \overline{\ell^0_{L\alpha}} i \not{D} \ell^0_{L\alpha} + \overline{e^0_{R\alpha}} i \not{D} e^0_{R\alpha}$
$\mathcal{L}_{\rm Higgs}$	$(D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi)$
$\mathcal{L}_{\rm Yukawa}$	$-Y^d_{\alpha\beta} \, \overline{q^0_{L\alpha}} \phi d^0_{R\beta} - Y^u_{\alpha\beta} \, \overline{q^0_{L\alpha}} \tilde{\phi} u^0_{R\beta} - Y^\ell_{\alpha\beta} \, \overline{\ell^0_{L\alpha}} \phi e^0_{R\beta} + \text{h.c.}$

In the SM, step **(3)** is is achieved through the scalar doublet field ϕ, or Higgs field. In the Higgs sector, the Lagrangian $\mathcal{L}_{\rm Higgs}$ contains the scalar potential $V(\phi)$ which has the general form

$$V(\phi) = \mu^2_\phi \, \phi^\dagger \phi + \frac{\lambda_\phi}{2} (\phi^\dagger \phi)^2 = \frac{\lambda_\phi}{2} \left(\phi^\dagger \phi + \frac{\mu^2_\phi}{\lambda_\phi} \right)^2 + \text{const.} \, . \tag{5}$$

The Higgs potential is responsible for the electroweak symmetry breaking $SU(2)_L \otimes U(1)_Y \to U(1)_Q$. This can be achieved spontaneously when the mass parameter μ_ϕ^2, in Eq. (5), becomes negative. In this scenario $\langle \phi^\dagger \phi \rangle = 0$ becomes a local maximum and the absolute minimum is shifted to the non-zero vacuum expectation value $\langle \phi^\dagger \phi \rangle \equiv v^2 = -2\mu_\phi^2/\lambda_\phi$. The Higgs field can be rewritten in a more convenient basis, making use of the gauge freedom, in which only the physical components (the ones associated with physical particles) are present. This is known as the unitary gauge and the scalar doublet takes the form

$$\phi = \begin{pmatrix} 0 \\ \dfrac{v+h}{\sqrt{2}} \end{pmatrix}, \quad \textbf{degrees of freedom}: \begin{cases} \phi^+ \text{ and } \mathrm{Im}\{\phi^0\} \text{ are the Goldstone bosons. ``Rotated away'';} \\ \\ \mathrm{Re}\{\phi^0\} \text{ was shifted, such that } h \text{ represents the true oscillations around the absolute minimum.} \end{cases} \tag{6}$$

In this basis it becomes clear that the gauge part of the kinetic term in the Higgs Lagrangian induces masses to some of the gauge bosons, i.e. to the ones associated with the broken generators,

$$(D_\mu \phi)^\dagger (D^\mu \phi) \sim m_W^2 W_\mu^+ W^{\mu -} + \frac{1}{2} m_Z^2 Z_\mu^0 Z^{\mu 0} + \cdots, \quad \text{with:} \begin{cases} m_W^2 = \dfrac{g^2 v^2}{4}, \quad m_Z^2 = \dfrac{g^2 v^2}{4 c_W^2}, \\ \\ m_A = 0 \quad \text{and} \quad m_G = 0. \end{cases} \tag{7}$$

Before closing this short overview on the SM building blocks, it is useful to do a simple consistency check and look at the degrees of freedom in the process of spontaneous symmetry breaking (SSB). We can restrict ourself to the $SU(2)_L \otimes U(1)_Y \to U(1)_Q$ sector. Before SSB, the theory consists of one complex scalar doublet field (four degrees of freedom) and four gauge bosons (two degrees of freedom each); there are $4 + 2 \times 4 = 12$ degrees of freedom. After the SSB, only $U(1)_Q$ remains as an explicit symmetry, i.e. only one generator leaves the vacuum invariant, so one would expect three Nambu-Goldstone bosons associated to the broken generators. Since we are working with a local gauge group, the Higgs mechanism allows these bosons to be absorbed as the longitudinal polarization of gauge bosons, $W^\pm$ and Z^0. So, in the end, we will have one real scalar field (one degree of freedom), three massive gauge bosons (three degrees of freedom each), and one massless gauge boson (the photon with two degrees of freedom). Summing up, after SSB there are $1 + 3 \times 3 + 2 = 12$ degrees of freedom, the same as in the unbroken phase.

Note that no field except for the Higgs has a mass term in the unbroken phase. The Higgs mechanism is responsible for the mass generation of fermions and gauge bosons, but not of its own mass!

2.4 The flavour structure of the SM

The origin of a non-trivial flavour structure in the SM is directly related with the presence of Yukawa interactions and gauge currents. The fermionic kinetic term is responsible for the weak charged currents (CC), weak neutral currents (NC) and for the electromagnetic neutral currents. They are given by

$$\textbf{Charged Current:} \quad \mathcal{L}_{\text{CC}} = \frac{g}{\sqrt{2}} \left(\overline{u_{L\alpha}^0} \gamma^\mu d_{L\alpha}^0 W_\mu^+ + \overline{e_{L\alpha}^0} \gamma^\mu \nu_{L\alpha}^0 W_\mu^- \right) + \text{h.c.}, \tag{8a}$$

$$\textbf{Neutral Current:} \quad \mathcal{L}_{\text{NC}} = e Q_f \overline{f^0} \gamma^\mu f^0 A_\mu + \frac{g}{c_W} \overline{f^0} \gamma^\mu \left(g_V^f - g_A^f \gamma_5 \right) f^0 Z_\mu, \tag{8b}$$

where

$$g_V^f = \frac{1}{2} T_3^f - s_W^2 Q_f, \quad g_A^f = \frac{1}{2} T_3^f, \tag{9}$$

are the vector (V) and axial (A) couplings of the the gauge boson Z^0 to the fermions, respectively. The letter f denotes any of the fermion fields. The charge of a fermion is denoted by Q_f, while T_3^f denotes the weak isospin associated with the left-handed fermion.

When a theory has several fields with the same quantum numbers (flavours) one is free to rewrite the Lagrangian in terms of new fields, obtained from the original ones by means of a unitary transformation which mixes them. Why only unitary transformations? In principle, one can mix particles with the same quantum numbers in 'any way' we want. However, by keeping it unitary we guarantee that the kinetic terms remain unaltered. This is important since having the kinetic Lagrangian with no cross terms, known as the canonical basis, allow us to easily identify our field content. We can define a set of transformations called weak basis transformations (WBTs) which are defined as transformations of the fermion fields which leave invariant the kinetic terms as well as the gauge interactions, i.e. they respect the gauge symmetry in the unbroken phase. The WBTs depend on the gauge theory that one is considering because, if there are more gauge interactions, then, in principle there will be less freedom to make WBTs. In the SM we define the WBTs as

$$\textbf{WBTs:}\quad \begin{cases} q_L^0 = W_L^q q_L' \,,\; u_R^0 = W_R^u u_R' \,,\; d_R^0 = W_R^d d_R' \,, \\ \ell_L^0 = W_L^\ell \ell_L' \,,\; e_R^0 = W_R^e e_R' \,, \end{cases} \longrightarrow \quad \begin{cases} Y_u' = W_L^\dagger Y_u W_R^u \,, \\ Y_d' = W_L^\dagger Y_d W_R^d \,, \\ Y_e' = W_L^{\ell\dagger} Y_e W_R^e \,. \end{cases} \tag{10}$$

where $W_L^{q,\ell}$ and $W_R^{u,d,e}$ are 3×3 unitary matrices acting in the flavour space. The transformed Yukawa matrices $Y'_{u,d,e}$ have the same physical content as the original ones. To see the usefulness of WBTs let us start from a general basis where the mass matrix $Y_{u,d,e}$ have 18 free parameters each (9 modulus and 9 phases). An arbitrary $n\times n$ complex matrix A can be diagonalized by a bi-unitary transformation as $U_L^\dagger A V_R = \text{diag}$. This is known as single value decomposition. Using this information we can pass from a general basis to the new basis

$$\begin{array}{lcl} \textbf{flavour basis I:} & & \textbf{flavour basis II:} \\ \begin{cases} Y_u = U_L^u \lambda_u V_R^{u\dagger} \\ Y_d = U_L^d \lambda_d V_R^{d\dagger} \\ Y_e = U_L^e \lambda_e V_R^{e\dagger} \end{cases} & \textbf{WBTs} & \begin{cases} Y_u' = V_{CKM}^\dagger \lambda_u \\ Y_d' = \lambda_d \\ Y_e' = \lambda_e \end{cases} \\ & W_L^q = U_L^d \,,\; W_R^u = V_R^u \,,\; W_R^d = V_R^d & \\ & W_L^\ell = U_L^e \,,\; W_R^e = V_R^e & \end{array} \,, \tag{11}$$

with $\lambda_u = \text{diag}(y_u, y_c, y_t)$, $\lambda_d = \text{diag}(y_d, y_s, y_b)$ and $\lambda_e = \text{diag}(y_e, y_\mu, y_\tau)$ the real and positive fermion Yukawas (defined from the fermion masses, i.e. $y_f = \sqrt{2} m_f / v$), and $V_{\text{CKM}} = U_L^{u\dagger} U_L^d$. This unitary matrix is the well known Cabbibo-Kobayashi-Maskawa (CKM) quark mixing matrix [25, 26]. As we shall see in a while, this matrix only has four degrees of freedom. Therefore, in the flavour basis II we only have $6(\text{masses}) + 4(\text{mixing}) = 10$ free parameters in the quark sector, mush less than in the general flavour basis I. Note that this is actually the minimal number of free parameters that one can have, since it is equal to the physical ones. Basis with less free parameters cannot be obtained by WBTs and they would have physical implications (correlations between physical observables).

The WBTs become much a more fundamental aspect of the model when $Y^{u,d,\ell} \to 0$. In this limit the WBTs given in Eq. (10) leave the whole Lagrangian invariant and therefore are promoted to symmetry generators of a global $U(3)^5$ symmetry

$$\mathcal{G}_{\text{global}} \equiv U(3)^5 = SU(3)_q^3 \times SU(3)_l^2 \times U(1)^5 \,, \tag{12}$$

where

$$SU(3)_q^3 = SU(3)_{q_L} \times SU(3)_{u_R} \times SU(3)_{d_R} \quad \text{and} \quad SU(3)_l^2 = SU(3)_{\ell_L} \times SU(3)_{e_R} \,. \tag{13}$$

In the presence of Yukawa terms only a reminiscent of the original global symmetry $\mathcal{G}_{\text{global}}$ remains unbroken. The easiest way to see which symmetry is left invariant is to look at the flavour basis II,

introduced in Eq. (11), in which the number of parameters is reduced to the physical ones. In this basis the only field transformations that leave the Lagrangian invariant are rephasing rotations, and the presence of the $V_{\rm CKM}$ matrix only allows one rotation in the quark sector. From this simple inspection we see that after the introduction of the Yukawa terms we are left with the residual symmetry

$$\mathcal{G}_{\rm global} \longrightarrow \mathcal{G}_{\rm global}^{\rm accidental} \equiv U(1)_B \times U(1)_e \times U(1)_\mu \times U(1)_\tau \,, \tag{14}$$

with, of course, the gauge $U(1)_Y$ symmetry unbroken. These are called accidental symmetries, they were not imposed in the SM construction but end up appearing as a consequence of renormalizability and perturbativity.

Looking at the WBTs as symmetry generators is actually very convenient in order to count the number of physical parameters present in the model. No mater which parameterization we choose for the SM flavour couplings $Y_{u,d,e}$, the number of physical parameters always remains unaltered. To learn how to count these parameters, let us first look at the charged lepton relevant flavour couplings $Y^e_{\alpha\beta}\overline{\ell^0_{L\alpha}}\phi e^0_{R\beta}$. Our goal is to find out how many of the 18 real parameters are actually physical. Now, if we look at the limit $Y^e \to 0$, we know that the Lagrangian will enjoy of a larger global symmetry, i.e. a $U(3)_{\ell_L} \times U(3)_{e_R}$ global symmetry. Another piece of information that is crucial is the residual symmetry of our model. Concerning the leptonic sector, as was seen above, we have the accidental $U(1)_e \times U(1)_\mu \times U(1)_\tau$. In other words, the presence of Y_e induces the breaking

$$\underbrace{\overbrace{U(1)_\phi}^{\text{Higgs}} \times \overbrace{U(3)_{\ell_L} \times U(3)_{e_R}}^{\text{Leptons}}}_{\text{1+9+9 generators}} \xrightarrow[Y_e]{} \underbrace{U(1)_e \times U(1)_\mu \times U(1)_\tau \times U(1)_Y}_{\text{1+1+1+1 generators}}, \tag{15}$$

$$\text{15 broken generators}$$

leading to the existence of 15 broken generators. We have included the Higgs and the hypercharge symmetries for completeness[1]. We can now use the broken generators to rotate Y^e into a "convenient" symmetry-breaking direction. These rotations are nothing more that the WBTs described in Eq. (11), resulting in three physical parameters, i.e. the charged lepton masses. The result found in this simple exercise is actually more general and can be stated as follows

$$\#\ \textbf{Physical parameters} \ = \ \#\ \textbf{Total parameters} - \#\ \textbf{Broken generators} \tag{16}$$

Let us apply this result to the quark sector, we have

$$\begin{aligned} &\#\ \textbf{Total parameters:}\ \overbrace{(9+9)}^{Y_d} + \overbrace{(9+9)}^{Y_u} = 36 \\ &\#\ \textbf{Broken generators:}\ \underbrace{3\times 9}_{U(3)^3} - \underbrace{1}_{U(1)_B} = 26 \end{aligned} \qquad \Longrightarrow \quad \#\ \textbf{Physical parameters:}\ 10\,. \tag{17}$$

Note that Eq. (14) is only true at the classical level since non-perturbative quantum effects break this down to just one abelian group $U(1)_{3B-L}$. However, this does not affect the parameter counting.

The Yukawa sector of the SM is responsible for the mass generation of the fermion species, after SSB. The fermion mass assignment in the SM is given by a Dirac mass term, $-m_f \bar{f} f = -m_f(\bar{f}_L f_R + \bar{f}_R f_L)$. Although it is invariant under $U(1)_Q$, the fermion mass term is not invariant under $SU(2)_L \otimes U(1)_Y$. Indeed, a fermion mass term is not a singlet under $SU(2)_L$, and, besides, the right- and left-handed components of f have different weak hypercharges. As a result, no pure fermionic mass terms

[1] Note that while in the SM these symmetries can be ignored in the process of counting broken generators, they play a crucial role in several extension of the SM.

can be constructed consistently with gauge invariant principles, as it was mention in the prevous section. In the SM fermion masses can arise from Yukawa interactions with the scalar Higgs doublet, i.e the Lagrangian part $\mathcal{L}_{\text{Yukawa}}$. Using the Higgs filed given in Eq. (6), one can see that the Yukawa Lagrangian splits into two parts, one relative to the fermion masses, $\mathcal{L}_{\text{mass}}$, and another corresponding to the interaction of the Higgs field with the fermions, $\mathcal{L}_{\text{hff}}$,

$$\textbf{Mass:} \quad -\mathcal{L}_{\text{mass}} = M^{e}_{\alpha\beta}\,\overline{e^{0}_{L\alpha}}e^{0}_{R\beta} + M^{u}_{\alpha\beta}\,\overline{u^{0}_{L\alpha}}u^{0}_{R\beta} + M^{d}_{\alpha\beta}\,\overline{d^{0}_{L\alpha}}d^{0}_{R\beta} + \text{h.c.}\,, \tag{18a}$$

$$\textbf{hff:} \quad -\mathcal{L}_{\text{hff}} = \frac{1}{\sqrt{2}}Y^{\ell}_{\alpha\beta}\,\overline{e^{0}_{L\alpha}}e^{0}_{R\beta}\,h + \frac{1}{\sqrt{2}}Y^{u}_{\alpha\beta}\,\overline{u^{0}_{L\alpha}}u^{0}_{R\beta}\,h + \frac{1}{\sqrt{2}}Y^{d}_{\alpha\beta}\,\overline{d^{0}_{L\alpha}}d^{0}_{R\beta}\,h + \text{h.c.}\,, \tag{18b}$$

with the fermion mass matrices given by

$$M^{f} = \frac{v}{\sqrt{2}}Y^{f}\,, \quad \text{with} \quad f = \{u, d, e\}\,. \tag{19}$$

At this stage it is worth pointing out that, in the SM, no renormalizable mass term for neutrinos can be constructed due to the absence of the right-handed fields ν_R. Also, a particular feature of the SM is to have the mass terms proportional to the Yukawa couplings, leading to the absence of flavour changing neutral currents (FCNC) in the scalar sector. Extensions beyond SM in general "struggle", i.e. need additional assumptions beyond new particles, in order to reproduce this alignment [27].

The Higgs mechanism breaks the $SU(2)_L$ group, which means that in the broken phase we are able to rotate the fields in the same $SU(2)_L$ multiplet through different unitary transformations. Therefore, we see from the new weak basis defined in Eq. (11) that we can redefine the field d_L as $d'_L = V_{\text{CKM}}d_L$ such that the mass matrices are both diagonal and charged current sector becomes

$$\mathcal{L}_{\text{CC}} = \frac{g}{\sqrt{2}}\left(\overline{u_{L\alpha}}\,(V_{\text{CKM}})_{\alpha\beta}\,\gamma^{\mu}d_{L\beta}W^{+}_{\mu} + \overline{e_{L\alpha}}\gamma^{\mu}\nu_{L\alpha}W^{-}_{\mu}\right) + \text{h.c.}\,, \tag{20}$$

with

$$V_{\text{CKM}} \equiv U^{u\dagger}_{L}U^{d}_{L} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}\,. \tag{21}$$

The unitary matrix present in the leptonic sector is the identity matrix since ν^{0}_{L} can be rotated freely through a unitary transformation, due to the absence of a mass term. Therefore, in the SM the only tree-level flavour-changing interactions are present in the charged currents. Since the matrix V_{CKM} is a 3×3 unitary matrix, it has 9 free parameters. However, the additional freedom

$$V_{\text{CKM}} \longrightarrow K^{\dagger}_{u}V_{\text{CKM}}K_{d}\,, \tag{22}$$

with $K_{u,d}$ phase diagonal matrices, reflecting the freedom in redefining the phases of the quarks in the mass basis, leads to 4 mixing parameters. Therefore, as stated before the weak basis in Eq. (11) has $4\,\text{mixing} + 6\,\text{masses} = 10$ parameters. This is known as the quark physical basis, since the number of free parameters coincides with the number of physical ones. Working in the mass eigenbasis, i.e. in the basis where the mass matrix of the fermions are real and positive, one can shift all the non-trivial flavour structure into the charged current sector. This is a very convenient basis to work in, since the fermion propagation gets quite simple. Still, we could opt to work in another basis at the cost of introducing extra complexity in the model.

In the SM CP violation shows up in the complex Yukawa couplings. If we CP conjugate a typical Yukawa term we get, see Table 2,

$$\mathcal{CP}\,\left(\overline{\psi_{L\alpha}}\phi\psi_{R\beta}\right)\mathcal{CP}^{\dagger} = \overline{\psi_{R\beta}}\phi^{\dagger}\psi_{L\alpha}\,. \tag{23}$$

We then see that by requesting CP invariance in the Yukawa sector we get

$$\mathcal{CP}\mathcal{L}_{\text{Yuk}}\mathcal{CP}^\dagger = \mathcal{L}_{\text{Yuk}} \quad \Rightarrow \quad Y_{\alpha\beta} = Y^*_{\alpha\beta}\,, \tag{24}$$

i.e. real Yukawa couplings are the necessary condition for CP-invariance. We can do the same exercise but now for the charged current Lagrangian, in the mass eigenbasis,

$$\mathcal{CP}\mathcal{L}_{\text{CC}}\mathcal{CP}^\dagger = \mathcal{L}_{\text{Yuk}} \quad \Rightarrow \quad V_{\alpha\beta} = V^*_{\alpha\beta}\,, \tag{25}$$

i.e. real CKM mixing matrix as the necessary condition. Therefore, the complex nature of the Yukawa couplings (or CKM mixing matrix) is the origin of CP violation in the SM. The above results are basis dependent. We know, that there are always phases that can be rotated away. So the question is whether we have a basis independent way of checking for CP violation. The answer is yes, the above conclusions can be formulated in a basis invariant way through the quantity [28]

$$\text{Tr}[H_u, H_d]^3 = 6i \sum_{\alpha,\beta=u,c,t,\ldots} \sum_{\alpha',\beta'=d,s,b,\ldots} = m^4_\alpha m^2_\beta m^4_{\alpha'} m^2_{\beta'}\, \text{Im} Q_{\alpha\alpha'\beta\beta'} \tag{26}$$

where

$$Q_{\alpha\alpha'\beta\beta'} \equiv V_{\alpha\alpha'} V_{\beta\beta'} V^*_{\alpha\beta'} V^*_{\beta\alpha'} \tag{27}$$

is the rephasing-invariant quartet. For three generations, the above invariant reads

$$\text{Tr}[H_u, H_d]^3 = 6i(m^2_t - m^2_c)(m^2_t - m^2_u)(m^2_c - m^2_u)(m^2_b - m^2_s)(m^2_b - m^2_d)(m^2_s - m^2_d)\, J\,, \tag{28}$$

with $J \equiv \text{Im} Q_{uscb} = \text{Im}[V_{us} V_{cb} V^*_{ub} V^*_{cs}]$ known as the Jarlskog invariant [29]. The CKM-mechanism is the origin of CP violation in the SM and lead to the nobel prize attribution in 2008 to Kobayashi and Maskawa who were the first to propose three flavours of quarks as the origin of CP violation [26].

Different parametrizations for the CKM mixing matrix can be used. We shall follow the standard procedure and use the Particle Data Group (PDG) parametrization [30]

$$\begin{aligned} V_{\text{CKM}} &= R_1(\theta_{23})\Gamma(\delta)R_2(\theta_{13})\Gamma(-\delta)R_3(\theta_{12}) \\ &= \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix} \end{aligned} \tag{29}$$

where $c_{ij} \equiv \cos\theta_{ij}$, $s_{ij} \equiv \sin\theta_{ij}$, $R(\theta_{ij})$ is the rotation in the plane $i-j$ and $\Gamma(\delta) = \text{diag}(1, 1, e^{i\delta})$. The three s_{ij} are the real mixing parameters and δ is the Kobayashi-Maskawa phase. While the range of this phase is $0 \leq \delta < 2\pi$, the measurements of $\mathcal{CP}$ violation in K decays force it to be in the range $0 < \delta < \pi$. From experiments we know that there exists a strong hierarchy on the mixing angles, i.e. $s_{13} \ll s_{23} \ll s_{12} \ll 1$. We can write the mixing angles as

$$\begin{aligned} s_{12} =& \lambda = \frac{|V_{us}|}{\sqrt{|V_{ud}|^2 + |V_{us}|^2}}\,, \quad s_{23} = A\lambda^2 = \lambda \left|\frac{V_{cb}}{V_{us}}\right|\,, \\ s_{13}e^{i\delta} =& V^*_{ub} = A\lambda^3(\rho + i\eta) = \frac{A\lambda^3(\bar\rho + i\bar\eta)\sqrt{1 - A^2\lambda^4}}{\sqrt{1-\lambda^2}[1 - A^2\lambda^4(\bar\rho + i\bar\eta)]}\,. \end{aligned} \tag{30}$$

With these relations we ensure that

$$\bar\rho + i\bar\eta = -\frac{V_{ud}V^*_{ub}}{V_{cd}V^*_{cb}} \tag{31}$$

is independent of any phase convention. The above expression allows us to express the CKM matrix in terms of: λ, A, $\bar\rho$ and $\bar\eta$. While the parametrization in term of these parameter is exact, it is common to

approximate this result for small λ. Up to fourth power corrections, we can expand the bar parameters as $\bar{\rho} = \rho(1-\lambda^2/2)$ and $\bar{\eta} = \eta(1-\lambda^2/2)$ known as Wolfenstein parametrization [31]

$$V_{\rm CKM} = \begin{pmatrix} 1-\lambda^2/2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\lambda^2/2 & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}(\lambda^4)\,. \tag{32}$$

The unitarity on the CKM matrix implies relations between its entries:

$$\begin{aligned} \textbf{Columns Orthogonality:} \quad & \sum_i V_{ij}V_{ik}^* = \delta_{jk}\,, \\ \textbf{Rows Orthogonality:} \quad & \sum_i V_{ij}V_{kj}^* = \delta_{ik}\,. \end{aligned} \tag{33}$$

The six vanishing combinations are sums of complex number, so that they can be represented as triangles in the complex plane. The most used triangle is given by

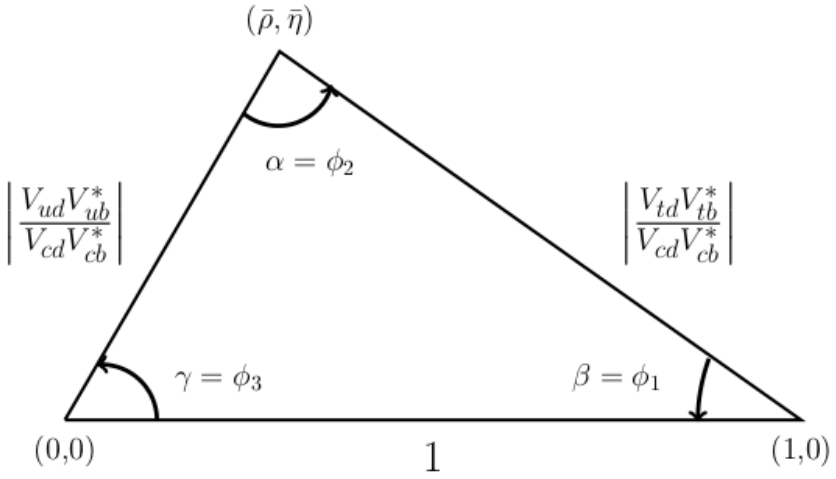


Fig. 1: Unitary triangle representation in the complex plane $\bar{\rho}, \bar{\eta}$

$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0\,,. \tag{34}$$

In Fig. 1 we have divided each side by the best-known value, i.e. $V_{cd}V_{cb}^*$. The angles of the unitary triangle are also represented in Fig. 1 and are given by

$$\beta = \phi_1 = \arg\left(-\frac{V_{cd}V_{cb}^*}{V_{td}V_{tb}^*}\right)\,, \quad \alpha = \phi_2 = \arg\left(-\frac{V_{td}V_{tb}^*}{V_{ud}V_{ub}^*}\right)\,, \quad \gamma = \phi_3 = \arg\left(-\frac{V_{ud}V_{ub}^*}{V_{cd}V_{cb}^*}\right)\,. \tag{35}$$

Measurements of CP-violating observables can constraint these angles and also the parameters $\bar{\eta}$, $\bar{\rho}$. Using the Wolfenstein parametrization as a give line, we can get simpler expressions for the unitary triangle angles

$$\begin{aligned} \beta =& \pi + \arg(V_{cd}V_{cb}^*) - \arg(V_{td}V_{tb}^*) \simeq -\arg(V_{td})\,, \\ \gamma =& \pi + \arg(V_{ud}V_{ub}^*) - \arg(V_{cd}V_{cb}^*) \simeq -\arg(V_{ub})\,. \end{aligned} \tag{36}$$

With the help of the unitary triangle where the d-quark is replaced by the s-quark, i.e.

$$V_{us}V_{ub}^* + V_{cs}V_{cb}^* + V_{ts}V_{tb}^* = 0\,, \tag{37}$$

we can define another angle

$$\beta_s = \arg\left(-\frac{V_{ts}V_{tb}^*}{V_{cs}V_{cb}^*}\right) = \pi + \arg(V_{ts}V_{tb}^*) - \arg(V_{cs}V_{cb}^*) \simeq \pi + \arg(V_{ts})\,. \tag{38}$$

This allow us to write the CKM mixing matrix up to $\mathcal{O}(\lambda^5)$ as

$$V_{\rm CKM} \simeq \begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}|e^{i\gamma} \\ -|V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}|e^{i\beta} & -|V_{ts}|e^{i\beta} & |V_{tb}| \end{pmatrix}\,. \tag{39}$$

The area of all triangles is the same and is given by half of the absolute value of the Jarlskog invariant, i..e Area$_\Delta = |J|/2$. The Jarlskog invariant in the parametrizations presented above take the form

$$J = \text{Im}\,[V_{ud}V_{cs}V_{us}^*V_{cd}^*] = \frac{1}{8}\sin(2\theta_{12})\sin(2\theta_{13})\sin(2\theta_{23})\sin\delta \simeq A^2\lambda^6\eta\,. \tag{40}$$

The absolute values of the CKM matrix can be found in the following processes:

- $|V_{ud}|$: β-decay $(A,Z) \to (A,Z+1) + e^- + \bar{\nu}_e$;
- $|V_{us}|$: K-decay $K^+ \to \pi^0 + \ell^+ + \nu_\ell$;
- $|V_{cd}|$: ν-production of c's $\nu_\ell + d \to \ell^- + c$;
- $|V_{cs}|$: charm decay $D^\pm \to K^0 + \ell^\pm + \nu_\ell$;
- $|V_{ub}|$: B-decay $b \to u + \ell^- + \bar{\nu}_\ell$;
- $|V_{cb}|$: B-decay $b \to c + \ell^- + \bar{\nu}_\ell$;
- $|V_{td}|$ and $|V_{ts}|$: Δm in $B^0 - \overline{B^0}$;
- $|V_{tb}|$: top decays.

The result of a global fit gives [30]

$$|V_{CKM}| = \begin{pmatrix} 0.97427 \pm 0.00014 & 0.22536 \pm 0.00061 & 0.00355 \pm 0.00015 \\ 0.22522 \pm 0.00061 & 0.97343 \pm 0.00015 & 0.0414 \pm 0.0012 \\ 0.00886^{+0.00033}_{-0.00032} & 0.0405^{0.0011}_{-0.0012} & 0.99914 \pm 0.00005 \end{pmatrix} \tag{41}$$

or in terms of the Wofenstein parameters

$$\lambda = 0.22537 \pm 0.00061\,, \quad A = 0.814^{+0.023}_{-0.024}\,, \quad \bar{\rho} = 0.117 \pm 0.021 \quad \text{and} \quad \bar{\eta} = 0.353 \pm 0.013\,. \tag{42}$$

The Jarlskog invariant is $J = (3.06^{+0.21}_{-0.20}) \times 10^{-5}$. The angles of the unitary triangle can be tested in B-decays:

- $\sin 2\beta$: $B_d^0 \to J/\Psi K_S$
- $\sin 2\alpha$: $B_d^0 \to \pi^+\pi^-$
- $\sin 2\gamma$: $B_s^0 \to D_S^\pm K^\mp$

2.5 GIM mechanism

We have learned that the structure of the SM is such that it ensures the absence of the tree level flavour changing neutral currents. Both neutral gauge boson and Higgs boson couplings are diagonal in the flavour mass eigenstate basis. Thus, the flavour changing neutral-current processes involving quarks are generated in higher orders in the electroweak interactions. Since they are strongly suppressed in Nature, it is interesting to discuss the predictions for them in the electroweak theory. For the quark sector, the generic examples of flavour changing neutral-current transitions are the reactions:

- $d\bar{s} \to \bar{d}s$ $(\Delta S = 2)$, $\quad b\bar{d} \to \bar{b}d$ $(\Delta B = 2)$;
- $s \to d\gamma$ $(\Delta S = 1)$, $\quad b \to s\gamma$ $(\Delta B = 1)$.

Such transitions are responsible for physical processes like $K^0 - \overline{K^0}$ and $B^0 - \overline{B^0}$ mixing, for radiative flavour changing decays of strange and bottom mesons and for decays like $K \to \pi e^+ e^-$ or $B \to K^* e^+ e^-$. On dimensional grounds, we then get the following estimate for the $\bar{s}d \to \bar{s}d$ transition

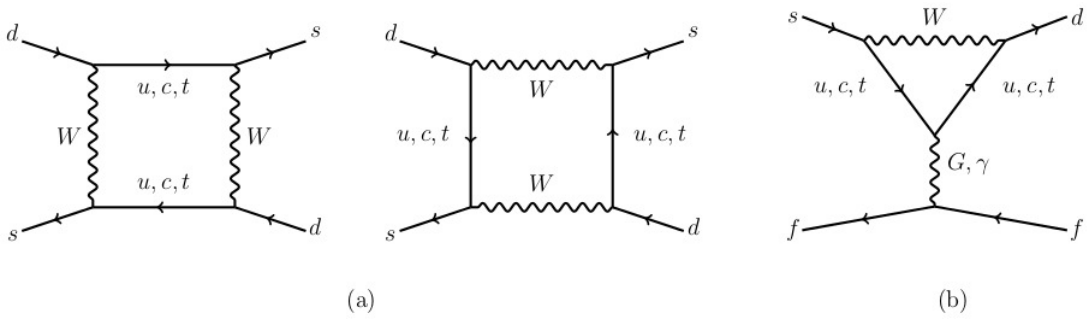


Fig. 2: In (a) $\Delta S = 2$ box diagrams. In (b) $\Delta S = 1$ penguin contribution.

amplitude, depicted in Fig. 2a , with double W-boson, u- and/or c-quark exchange (the contribution from the top quark exchange is strongly suppressed by its very small mixing with the first two generations of quarks):

$$\begin{aligned} A \sim & \left(\frac{e}{\sqrt{2}s_W}\right)^4 \frac{1}{M_W^2} \sum_{i,j=u,c} V_{is}^* V_{id} V_{js}^* V_{jd} \left[1 + \mathcal{O}\left(\frac{m_{q_i}^2}{M_W^2}, \frac{m_{q_j}^2}{M_W^2}\right)\right] \\ \sim & \alpha G_F \left[(V_{ts}^* V_{td})^2 + \mathcal{O}\left(\sum_{i,j=u,c} V_{is}^* V_{id} V_{js}^* V_{jd} \frac{m_q^2}{M_W^2}\right)\right] \end{aligned} \tag{43}$$

In the last step we have used the CKM unitarity condition: $\sum_{i,j=u,c} V_{is}^* V_{id} = -V_{ts}^* V_{td}$. We then see that the leading term is suppressed by very small CKM angles as the double top quark exchange contribution. The remaining terms, which are proportional to larger CKM angles, are in turn suppressed by light quark masses.

Such a mechanism of suppression of the flavour changing neutral-current amplitudes is known as the Glashow-Iliopoulos-Maiani (GIM) mechanism [32]. The strong suppression of the flavour changing neutral-current transitions is indeed a SM prediction. However, this follows not only from the structure of the theory but also depends on the empirical pattern of the quark masses and mixing angles. Therefore, from the SM point of view, the successful predictions for the flavour changing neutral-current processes are rather accidental.

Let us now look at the $\Delta F = 1$ transitions at the qualitative level. At one-loop, they receive contributions from box diagrams and also from the so-called penguin diagrams like in Fig. 2b. The corresponding amplitude goes as

$$A \sim \alpha G_F \sum_{i,j=u,c} V_{id}^* V_{is} \ln \frac{m_{q_i}^2}{M_W^2} + \mathcal{O}(V_{td}^* V_{ts}) = \alpha G_F V_{ud}^* V_{us} \ln \frac{m_u^2}{m_c^2} + \mathcal{O}(V_{td}^* V_{ts})\,. \tag{44}$$

Note that the dimensionless coefficient of the first term contains logarithms of light quark masses. Since the masses of the up and charm quarks are quite different, there is no additional suppression except for the usual one in this case (unlike the previously considered box diagrams). We can then say that the GIM mechanism is power-like in the case of box diagrams, but only logarithmic in the case of certain penguin diagrams.

3 Effective theories and their use in flavour physics

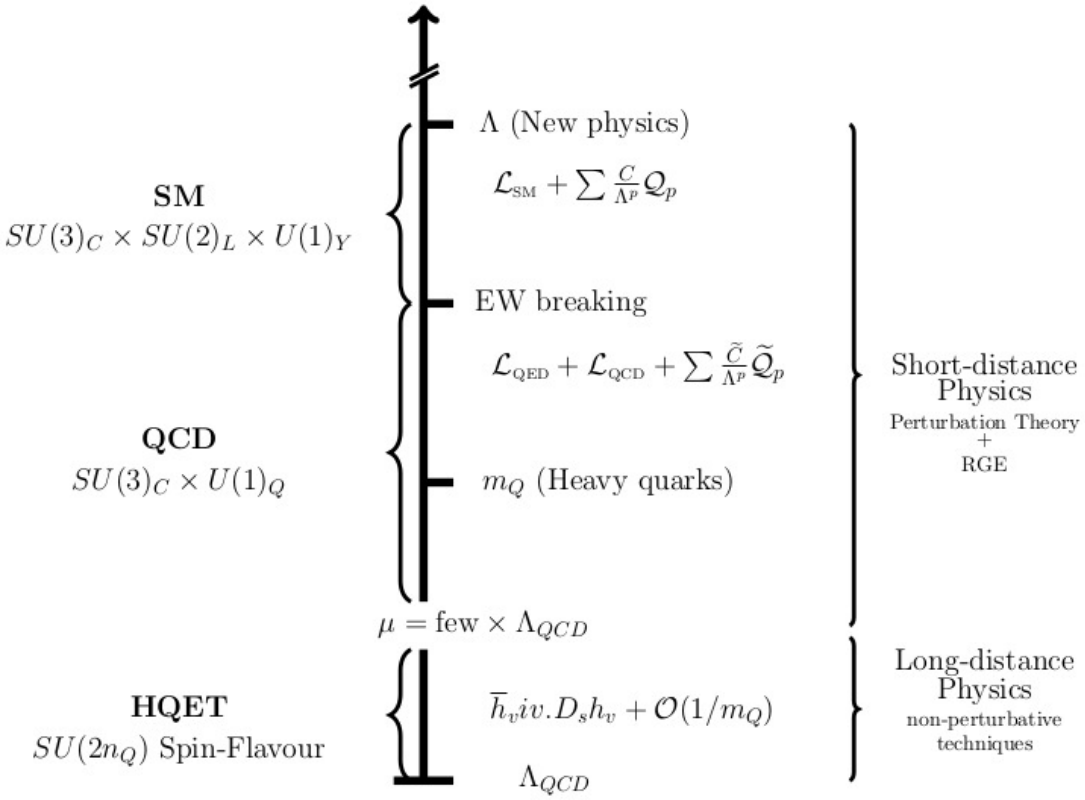


Fig. 3: General schematic idea behind effective field theories

Effective field theory formalism is a very powerful tool when several scales are present in a quantum field theory. The principle in effective field theories is to just include the appropriate degrees of freedom to describe physical phenomena occurring at a given scale. By integrating out degrees of freedom at shorter distances we try to simplify the model at longer distances. This approach works best when there is a large separation between length scale of interest and the length scale of the underlying dynamics. Figure 3 summarizes the general philosophy behind this approach. We summarize the effective field theory formalism in three simple steps [23]:

- **Step 1:** Choose a cutoff scale $\Lambda \lesssim M$ (with M some fundamental scale) and divide the field into high- and low-frequency modes, i.e.

$$\phi = \underbrace{\phi_H}_{\substack{\text{Fourier modes} \\ \omega > \Lambda}} + \underbrace{\phi_L}_{\substack{\text{Fourier modes} \\ \omega < \Lambda}} . \tag{45}$$

  The component ϕ_L describes the low-energy physics through the correlation functions

$$\langle 0|T\{\phi_L(x_1)\cdots\phi_L(x_n)\}|0\rangle = \frac{1}{Z[0]}\left(-i\frac{\delta}{\delta J_L(x_1)}\right)\cdots\left(-i\frac{\delta}{\delta J_L(x_n)}\right) Z[j_L]\Big|_{J_L=0} , \tag{46}$$

  where the generating functional is

$$Z[J_L] = \int \mathcal{D}\phi_L \mathcal{D}\phi_H \, e^{i\mathcal{S}(\phi_L,\phi_H)+i\int d^D x J_L(x)\phi_L(x)} \quad \text{and} \quad \mathcal{S}(\phi_L,\phi_R) = \int d^D x \mathcal{L}(x) . \tag{47}$$

  We have used D for the space-time dimension and only the external source of the low-frequency modes is relevant for the correlation functions computed at low energy.

- **Step 2:** Integrate out the high-frequency modes below the scale Λ, i.e.

$$Z[J_L] \equiv \underbrace{\int \mathcal{D}\phi_L \, e^{i\mathcal{S}_\Lambda(\phi_L)+i\int d^D x J_L(x)\phi_L(x)}}_{\text{No } \phi_H \text{ dependence}} \quad \text{and} \quad e^{iS_\Lambda(\phi_L)} = \int \mathcal{D}\phi_H e^{i\mathcal{S}(\phi_L,\phi_H)} \,. \tag{48}$$

  The action $S_\Lambda(\phi_L)$ is known as "Wilsonian effective action", which is non-local on scale $\Delta x^\mu \sim 1/\Lambda$ and depends on the choice made for the cutoff scale Λ.

- **Step 3:** Expand the non-local action in terms of local operators composed of light fields, which is known as operator-product expansion (OPE). This expansion is possible in the low-energy regime, i.e. $E \ll \Lambda$, and leads to

$$S_\Lambda(\phi_L) = \int d^D x \mathcal{L}^{\text{eff}}_\Lambda(x) \,, \quad \text{with} \quad \underbrace{\mathcal{L}^{\text{eff}}_\Lambda(x) = \sum_i \overbrace{C_i}^{\substack{\text{Wilson}\\ \text{coeff.}}} \overbrace{\mathcal{Q}_i(\phi_L(x))}^{\substack{\text{local}\\ \text{operator}}}}_{\text{Effective Lagrangian}} \tag{49}$$

The procedure described above is quite general and powerful, allowing us to obtain the Lagrangian relevant for a given scale. However, the effective Lagrangian is a sum of infinite operators which would naively destroy the predictability of the effective theory. In order to understand why this is not the case one can use the remarkably simple and powerful "naive dimensional analysis" (NDA) approach:

$$\begin{array}{c}\textbf{NDA:}\\ (c=\hbar=1)\end{array} \quad \begin{cases} [m]=[E]=[p]=[x^{-1}]=[t^{-1}]=1 \\ \\ \text{Assuming } [C_i] = -\gamma_i \end{cases} \quad \text{then} \quad C_i = g_i M^{-\gamma_i} \,. \tag{50}$$

The coupling g_i is dimensionless and form "naturalness" $\mathcal{O}(1)$, while M is the fundamental energy scale of the theory. Taken for simplicity the effective Lagrangian dimensionless, the effective operator $\mathcal{Q}_i$ scales for $E \ll \Lambda < M$ as

$$g_i \left(\frac{E}{M}\right)^{\gamma_i} = \begin{cases} \mathcal{O}(1) & \text{if } \gamma_i = 0 \\ \ll 1 & \text{if } \gamma_i > 0 \\ \gg 1 & \text{if } \gamma_i < 0 \end{cases} \tag{51}$$

This tell us that only the couplings that have $\gamma_i < 0$ are relevant. Therefore, given a precision goal we can truncate the series in $\mathcal{L}_\Lambda$ in a given order in E/M. This implies a finite number of operators, which brings back the predictability of the effective theory. The dimension γ_i can change due to interactions, this is known as anomalous dimension. We can be more formal and require the action to be dimensionless. In this case if $\delta_i = [\mathcal{O}_i]$ the coefficient dimension is $\gamma_i = \delta_i - D$. We summarize the operator relevance classification in Table 5.

As a final comment note that while most of the time ϕ_H is identified with a heavy particle, the method presented above is much more general. As opposed to integrate out some heavy particle, we can work on a scenario where only light particles are present. In this case we can lower the cutoff scale Λ by a small amount $\Lambda - \delta\Lambda$ and integrate out high frequencies of the light particle. This implies that the operators $\mathcal{O}_i(\phi_L)$ will remain the same, as no contribution from extra particles are present. And the effects of lowering the cutoff scale must enter into the effective couplings $C_i(\Lambda)$. This approach gives an intuitive understanding of the running of the coupling constants.

3.1 Weak currents and OPE

Hadrons can decay through weak interaction mediation, between their quark constituents. The typical binding energy of quarks in hadrons is $\mathcal{O}(1\,\text{GeV})$, much below the weak scale $\mathcal{O}(M_{W,Z})$. The idea

Table 5: Classification of operators based on their dimension.

Dimension	**Importance for $E \to 0$**	**Terminology**
$\delta_i < D$, $\gamma_i < 0$	grows	Relevant operators (super-renormalizable) • usually unimportant; • protected by symmetries
$\delta_i = D$, $\gamma_i = 0$	constant	Marginal operators (renormalizable) • renormalizable QFT
$\delta_i > D$, $\gamma_i > 0$	falls	Irrelevant operators (non-renormalizable) • the most important (relevant) • sensitive to fundamental scale

behind the OPE treatment is to start from short-distance dynamics and refine it step-by-step with non-perturbative corrections. Let us look at the part of generating functional containing the W boson [6], i.e.

$$Z_{\rm W} \sim \int [dW^+][dW^-] \text{Exp}\left(i \int d^4x \mathcal{L}_W\right) , \tag{52}$$

with

$$\begin{aligned} \mathcal{L}_{\rm W} = &- \frac{1}{2}\left(\partial_\mu W^+_\nu - \partial_\nu W^+_\mu\right)\left(\partial^\mu W^{-\nu} - \partial^\nu W^{-\mu}\right) + M_W^2 W^+_\mu W^{-\mu} \\ &+ \frac{g_2}{2\sqrt{2}}(J^+_\mu W^{+\mu} + J^-_\mu W^{-\mu}) \end{aligned} \tag{53}$$

the Lagrangian density containing the kinetic terms of the W boson and its interactions with charged currents. These interactions can be extracted from Eq. (20). Since we are not interested in W as external sources, we have omitted gauge self-interactions. Following the usual procedure in QFT, we can perform a Gaussian functional integration which leads us to a non-local action for quarks

$$\mathcal{S}_{\rm nl} = \int d^4x \mathcal{L}_{kin} - \frac{g_2^2}{8} \int d^4x d^4y \, J^-_\mu(x) \Delta^{\mu\nu}(x,y) J^+_\nu(y) \, , \tag{54}$$

where $\Delta^{\mu\nu}(x,y)$ is the W boson propagator. In the unitary gauge it reads

$$\Delta^{\mu\nu}(x,y) = \int \frac{d^4k}{(2\pi)^4} \Delta_{\mu\nu}(k) e^{-ik(x-y)} \, , \quad \Delta^{\mu\nu}(k) = \frac{-1}{k^2 - M_W^2}\left(g_{\mu\nu} - \frac{k_\mu k_\nu}{M_W^2}\right) . \tag{55}$$

The idea now is to formally expand in $1/M_W^2$ powers the propagator, which allows us to get a local action. To lowest order the propagator becomes

$$\Delta^{\mu\nu}(x,y) \simeq \frac{g^{\mu\nu}}{M_W^2} \delta^{(4)}(x-y) \, , \tag{56}$$

which in turns lead to the effective Hamiltonian

$$\mathcal{H}_{\rm eff} = -\frac{G_F}{\sqrt{2}} J^-_\mu J^{+\mu}(x) = -\frac{G_F}{\sqrt{2}} V^*_{\alpha\beta} V_{\alpha'\beta'} (\overline{d_\alpha} u_\beta)_{V-A} (\overline{d_{\alpha'}} u_{\beta'})_{V-A} \, . \tag{57}$$

We have adopt the notation $(\bar\psi\chi)_{V\mp A} \equiv \bar\psi\gamma^\mu(1 \mp \gamma_5)\chi$. This simple example introduces the main idea behind OPE, as already mentioned in the previous section. The above computation is nothing more than the usual 'integrating out' in effective theories. While we have used a path integral approach, the computation done is equivalent to the expansion of the W boson propagator in the amplitude matrix element, obtained from the usual Feynman rules approach (Fig. 4).

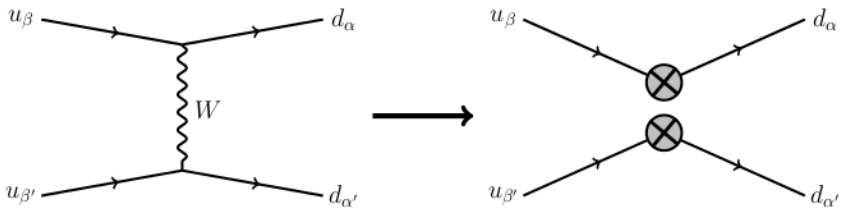


Fig. 4: Diagrammatic representation of the new local operators obtain from OPE formalism

Therefore, in general the OPE allows us to write an effective Hamiltonian of the form

$$\mathcal{H}_{\text{eff}} = \frac{G_F}{\sqrt{2}} \sum_i \lambda^i_{\text{CKM}} C_i(\mu) \mathcal{Q}_i \,, \tag{58}$$

where λ^i_{CKM} contains CKM factors (1 for semi-leptonic operators, 2 for quark operators), $C_i(\mu)$ are the Wilson coefficients and $\mathcal{Q}_i$ is a local operator governing the process in question. The coefficients $C_i(\mu)$ are weights of the operators $\mathcal{Q}_i$ on the effective Hamiltonian, i.e. they describe the strength with which a given operator contributes to the Hamiltonian. These are scale dependent couplings and can be calculated using perturbative methods (as long the scale μ is not too small). The operators $\mathcal{O}_i$ are the leading terms in the short-distance expansion described above; in the cases we are interested in, these will correspond to four-fermion operators. Therefore, at short distances we see processes mediated by heavy particles as point-like interactions.

We are interested in evaluating decay amplitude for a given type of meson P. With the help of the effective Hamiltonian this can be done quite 'easily' using

$$A(P \to F) = \langle F|\mathcal{H}_{eff}|P\rangle = \frac{G_F}{\sqrt{2}} \sum_i \lambda^i_{\text{CKM}} C_i(\mu) \langle F|\mathcal{Q}_i(\mu)|P\rangle \,, \tag{59}$$

where F denotes the final state, i.e we are looking at $P \to F$. The matrix element $\langle F|\mathcal{Q}_i(\mu)|P\rangle$ is evaluated at the renormalization scale μ and is the step that in general requires non-perturbative methods.

Equation (59) and Fig. 5 compiles the essence of the OPE method which allow the calculation of an amplitude $A(P \to F)$ to be factorize into two contributions:

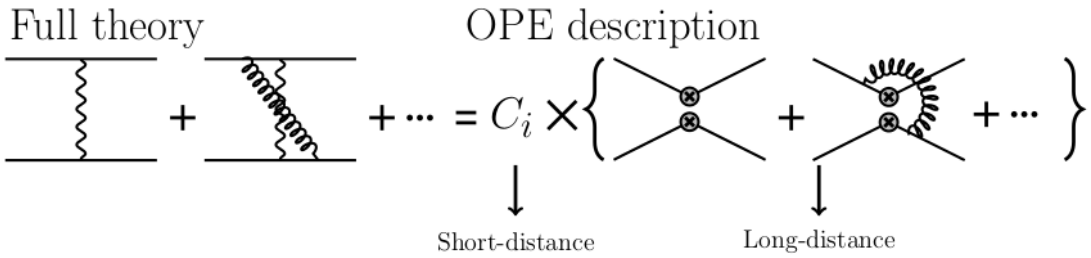


Fig. 5: Typical full theory description vs. OPE description

Short-distance effects

The computation of short-distance effects, or perturbative calculation, are all contained in the Wilson coefficients $C_i(\mu)$. These coefficients will include the contributions from integrating out the heavy particles such as top quarks, gauge bosons W and Z, and any new heavy field present in SM extensions. All effects of QCD interactions above the factorization scale μ are contained in these coefficients. $C_i(\mu)$ are independent of external states. This means that they are always the same no matter we consider the physical amplitudes where quarks are bound inside mesons, or any other unphysical amplitude with on-shell or off-shell quarks in the external lines.

Long-distance effects
The computation of long-distance effects is present in the calculation of the matrix element $\langle \mathcal{Q}_i(\mu) \rangle$. This means that all low-energy contributions below the factorization scale μ are encoded in the matrix element. The task is then to evaluate local operators between hadron states. This is the hardest task to do in the OPE treatment, since it requires in general a non-perturbative analysis.

As we saw, the most difficult aspect of OPE is the non-perturbative computation of $\langle \mathcal{Q}_i(\mu) \rangle$. Still the method offers a considerably simplified approach to the full amplitude computation. Next we shall illustrate the OPE in the context of $K^0 \to \pi^+\pi^-$ decay. We are, therefore, interested in the transition

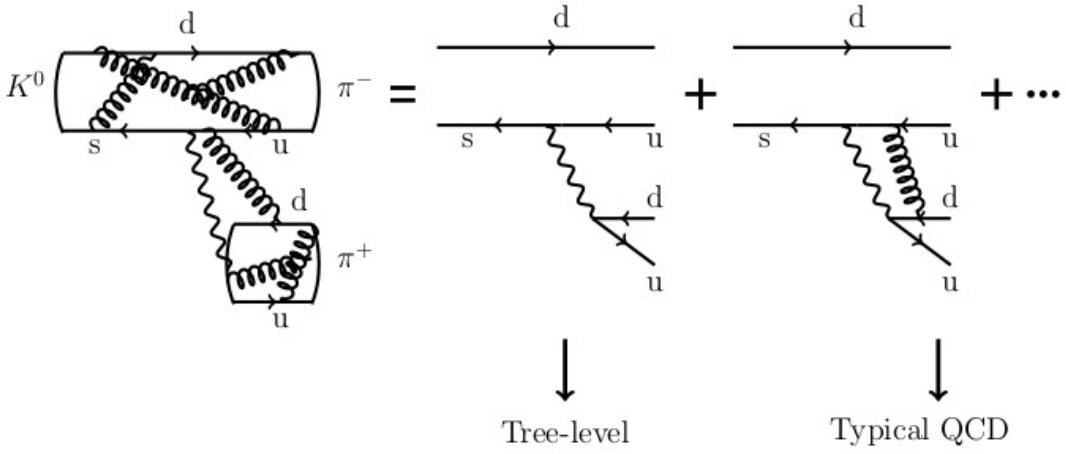


Fig. 6: General representation of $K^0 \to \pi^+\pi^-$ decay. The two diagrams on the right are the typical leading contributions.

$s \to uud$ as shown in Fig. 6. A convenient choice is to take all the light quarks to be massless and with the same off-shell momentum p. The Wilson coefficients $C_i(\mu)$ can then be found in perturbation theory from the 3 simple steps:

(1) Compute the amplitude (A_{full}) of the process in the full theory, i.e. in the presence of the W propagator, for arbitrary external states
(2) Compute the matrix element $\langle Q_i(\mu) \rangle$ with the same treatment for external states
(3) Compute $C_i(\mu)$ from the relation $A_{full} = A_{eff} = \frac{G_F}{\sqrt{2}} \sum_i \lambda^i_{CKM} C_i(\mu) \langle Q_i(\mu) \rangle$; this is known as matching of the full theory onto the effective one

Note that the choice of momenta leads to a gauge dependent amplitude. However, this cancels out with the gauge dependence from $\langle Q_i(\mu) \rangle$ such that $C_i(\mu)$ is physical. To order $\mathcal{O}(\alpha_S)$ we have four diagrams contributing: 1 with just W propagator; 1 ($\times$ 3 combinations) with W and gluon. Without QCD corrections we get the effective dimension 6 operator

$$\mathcal{Q}_2 = (\bar{s}_i u_i)_{V-A} (\bar{u}_j d_j)_{V-A} \,, \tag{60}$$

with i, j color indices (the notation $\mathcal{Q}_2$ is for historical reasons.). When QCD corrections are taken into account we at at order $\mathcal{O}(\alpha_S)$ the effective operator

$$\mathcal{Q}_1 = (\bar{s}_i u_j)_{V-A} (\bar{u}_j d_i)_{V-A} \,, \tag{61}$$

which resembles $\mathcal{Q}_2$ apart from the different color structure (see Fig. 7). This structure is obtained with the help of the $SU(N)$ Gell-Mann matrices identity

$$(\bar{s}_i T^a_{ik} u_k)(\bar{u}_j T^a_{jl} d_l) = -\frac{1}{2N} (\bar{s}_i u_i)(\bar{u}_j d_j) + \frac{1}{2} (\bar{s}_i u_j)(\bar{u}_j d_i) \,. \tag{62}$$

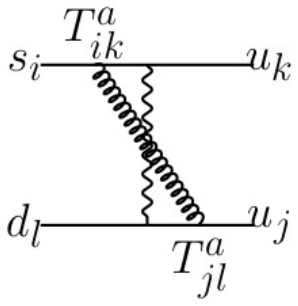


Fig. 7: Colour structure of typical QCD correction

Gluonic corrections to the matrix element of the original operator $\mathcal{Q}_2$ involve not just contributions from itself but additional structure from $\mathcal{Q}_1$. We say that the operators $\mathcal{Q}_1$ and $\mathcal{Q}_2$ mix under renormalization. Therefore, a convenient basis for the above operators is

$$\mathcal{Q}_\pm = \frac{\mathcal{Q}_2 \pm \mathcal{Q}_1}{2}\,, \quad C_\pm = C_2 \pm C_1\,, \tag{63}$$

where the renormalization of $+$ and $-$ are independent. We can then evaluate the full amplitude, which gives

$$-iA_{full} = -i\frac{G_F}{\sqrt{2}}V_{us}^*V_{ud}\left[\left(1+\gamma_+\alpha_s\ln\frac{M_W^2}{-p^2}\right)S_+ + \left(1+\gamma_-\alpha_s\ln\frac{M_W^2}{-p^2}\right)S_-\right], \tag{64}$$

where $S_\pm$ is the tree-level matrix elements of $\mathcal{Q}_\pm$ and $\gamma_\pm$ some numbers to be specify. This ends our first step. Next, we compute the matrix elements in the effective theory, which is given by

$$-i\langle\mathcal{Q}_\pm\rangle = -i\frac{G_F}{\sqrt{2}}V_{us}^*V_{ud}\left[1+\gamma_\pm\alpha_s\left(\frac{1}{\epsilon}+\ln\frac{\mu^2}{-p^2}\right)\right]S_\pm\,. \tag{65}$$

The last step is matching. From Eq. (64) and Eq. (65) one easily reads the Wilson coefficient to be

$$C_\pm = 1+\gamma_\pm\alpha_s\ln\frac{M_W^2}{\mu^2}\,. \tag{66}$$

A note of caution is in order. In the computation of the amplitude we did not perform any quark field renormalization. However, the renormalization in the effective theory can be explicitly seen in Eq. (65). Having divergent Wilson coefficients would be a clearly signal of inconsistency. Therefore, the above result was obtained after a renormalization on $\langle Q_\pm\rangle$ and using the MS scheme [6]. The presence of this divergence in Eq. (65) is directly linked to the $\ln M_W$ dependence of the decay amplitude in the full theory, which diverges in the limit $M_W \to \infty$.

Summing up, the effective Hamiltonian describing $K^0 \to \pi^+\pi^-$ decay is given by

$$\mathcal{H}_{\text{eff}} = \frac{G_F}{\sqrt{2}}V_{us}^*V_{ud}\left(C_+(\mu)\mathcal{Q}_+ + C_-(\mu)\mathcal{Q}_-\right) \tag{67}$$

up to $\mathcal{O}(\alpha_s\log)$ and with $C_\pm$ given by Eq. (66). In obtaining the decay amplitude from Eq. (67), the matrix elements $\langle 2\pi|\mathcal{Q}_\pm|K\rangle$ have to be taken, normalized at an appropriated scale μ. A typical scale for K decays is $\mu \simeq 1\,\text{GeV} \ll M_W$. Going beyond leading logarithmic approximation $\mathcal{O}(\alpha_s\log)$ makes the Wilson coefficients and matrix elements scheme dependent. This scheme dependence is unphysical and cancels out in the product of Wilson coefficient and matrix elements, as long as both quantities are evaluated with the same scheme.

In the example above we have whitenessed in first hand the OPE factorization. Schematically, its has the following structure

$$\left(1+\alpha_S\gamma_\pm\ln\frac{M_W^2}{-p^2}\right)\to\left(1+\alpha_S\gamma_\pm\ln\frac{M_W^2}{\mu^2}\right)\left(1+\alpha_S\gamma_\pm\ln\frac{\mu^2}{-p^2}\right), \tag{68}$$

which is achieved from the splitting of the logarithm into the sum of two terms. From the integration over virtual moment point of view this splitting reads

$$\int_{-p^2}^{M_W^2}\frac{dk^2}{k^2}=\underbrace{\int_{\mu^2}^{M_W^2}\frac{dk^2}{k^2}}_{\substack{\text{Short-distance effects}\\ \text{or}\\ \text{large virtual momenta}}}+\underbrace{\int_{-p^2}^{\mu^2}\frac{dk^2}{k^2}}_{\substack{\text{Long-distance effects}\\ \text{or}\\ \text{low virtual momenta}}}. \tag{69}$$

At this stage it is important to have a closer look to the Wilson coefficients found above. We can rewrite them, for convenience, as

$$C_\pm=1+\frac{\gamma_\pm(\alpha_S)}{2}\ln\frac{\mu^2}{M_W^2},\quad\text{with}\quad\gamma_\pm(\alpha_S)=\frac{\alpha_S(\mu)}{4\pi}\gamma_\pm^{(0)},\quad\text{and}\quad\gamma_\pm^{(0)}=\begin{cases}4\\-8\end{cases}. \tag{70}$$

The factor multiplying the logarithm is $\mathcal{O}(1/10)$ for $\mu=1\,\text{GeV}$ and therefore sizeable for perturbation theory; the logarithm itself is large $\mathcal{O}(10)$ making perturbation theory to fail. We then have the scenario where the coupling constant is small, but we have large logarithms. This is actually a common situation in QFTs. The naive perturbation done in terms of the coupling constant is no longer enough, and we must resum the terms $(\alpha_S\ln\mu/M_W)^n$ to all orders n. This procedure reorganizes the pertubation series by solving the renormalization group equation (RGE) for the Wilson coefficients. The RGE for the Wilson coefficients follows from the fact that the unrenormalized coefficients $C_\pm^{(0)}=Z_cC_\pm$ are μ independent. This then leads us to

$$\frac{d}{d\ln\mu}C_\pm(\mu)=\gamma_\pm(\alpha_S)C_\pm(\mu)\quad\text{with}\quad\gamma_\pm=-Z_c^{-1}\frac{d}{d\ln\mu}Z_c\,. \tag{71}$$

The parameters $\gamma_\pm(\alpha_S)$ are also known as anomalous dimension of $C_\pm$. The Wilson coefficients are dimensionless numbers in the usual sense. However, because of the presence of the scale M_W in the logarithm, these coefficients will depend on the energy scale μ. Therefore, $\gamma_\pm(\alpha_S)$ are scaling dimensions, measuring the rate of change of these coefficients with a changing scale μ. In general, when not working in the diagonal basis, these scaling dimensions are matrices mixing all Wilson coefficients. Using the RGE for the coupling constant

$$\frac{d\alpha_S}{d\ln\mu}=-2\beta_0\frac{\alpha_S^2}{4\pi}, \tag{72}$$

we can solve Eq. (71)

$$C_\pm(\mu)=\left[\frac{\alpha_S(M_W)}{\alpha_S(\mu)}\right]^{\gamma_\pm^{(0)}/2\beta_0}C_\pm(M_W)=\left[\frac{1}{1+\beta_0(\alpha_S(\mu)/4\pi)\ln(M_W^2/\mu^2)}\right]^{\gamma_\pm^{(0)}/2\beta_0}, \tag{73}$$

where we have used the condition $C_\pm(M_W)=1$, since no large logarithms should be present at $\mu=M_W$. The expression above contains the logarithmic corrections $\alpha_S\ln M_W/\mu$ to all orders in α_S. This shows the general result that renormalization group method allows us to go beyond the naive perturbation theory.

Two final remarks are in order. This approach can be generalized to go from M_W down to m_c, for example. Then we can do this by steps, first evolving down to the scale m_b and then see the theory

below this scale as an effective theory where the b quark has been integrated out. One should satisfy the continuity of the running coupling at the threshold, also known as threshold effects. These effects should be, in general, taken in consideration in the running. The second important effect is the generation of QCD penguin operators.

3.2 Effective Hamiltonians: Some examples

In this section we summarize the Standard Model operator basis for FCNC processes, which is useful when computing quantities based on the OPE formalism. We use the notation $q = u, d, s, c, b$. The loop functions appearing in the Wilson coefficients are given by

$$\begin{aligned} \tilde{E}_0(x) =& -\frac{7}{12} + \mathcal{O}(1/x) \\ f(x) =& \frac{x}{2} + \frac{4}{3}\ln x - \frac{125}{36} + \mathcal{O}(1/x) \\ g(x) =& -\frac{x}{2} - \frac{3}{2}\ln x + \mathcal{O}(1/x). \end{aligned} \tag{74}$$

- **Current-current operators**

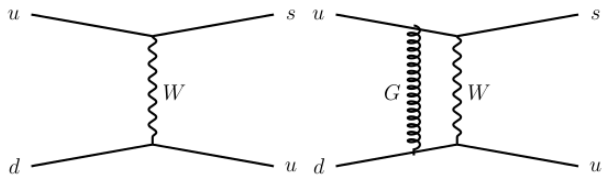


Fig. 8: Tree-level contribution and typical QCD correction topologies

$$\begin{aligned} \mathcal{Q}_1^p =& (\bar{s}_i p^i)_{V-A}(\bar{p}_j b^j)_{V-A}\,, \quad C_1(M_W) = 1 - \frac{11}{6}\frac{\alpha_s(m_W)}{4\pi} \\ \mathcal{Q}_2^p =& (\bar{s}_i p^j)_{V-A}(\bar{p}_j b^i)_{V-A}\,, \quad C_2(M_W) = \frac{11}{2}\frac{\alpha_s(m_W)}{4\pi} \end{aligned} \tag{75}$$

- **QCD Penguin operators:**

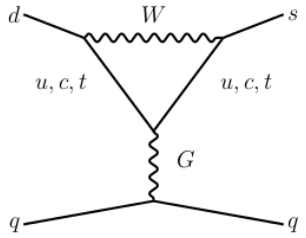


Fig. 9: QCD penguin topology

$$\begin{aligned} \mathcal{Q}_{3(5)} =& (\bar{s}_i b^i)_{V-A}\sum_q (\bar{q}_j q^j)_{V\mp A}\,, \quad C_{3(5)} = -\frac{1}{6}\tilde{E}_0\left(\frac{m_t^2}{m_W^2}\right)\frac{\alpha_s(m_W)}{4\pi} \\ \mathcal{Q}_{4(6)} =& (\bar{s}_i b^j)_{V-A}\sum_q (\bar{q}_j q^i)_{V\mp A}\,, \quad C_{4(6)} = \frac{1}{2}\tilde{E}_0\left(\frac{m_t^2}{m_W^2}\right)\frac{\alpha_s(m_W)}{4\pi} \end{aligned} \tag{76}$$

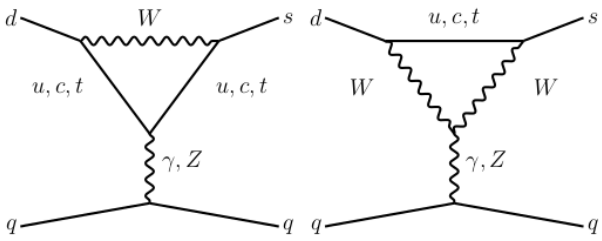


Fig. 10: Electroweak penguin topologies

- **Electroweak Penguin operators:**

$$\begin{aligned}\mathcal{Q}_{7(9)} =&(\overline{s}_i b^i)_{V-A} \sum_q \frac{3}{2} Q_q (\overline{q}_j q^j)_{V\pm A}\,,\\ C_7 = f\left(\frac{m_t^2}{m_W^2}\right)\frac{\alpha(m_W)}{6\pi}\,,&\quad C_9 = \left[f\left(\frac{m_t^2}{m_W^2}\right) + \frac{1}{s_W^2} g\left(\frac{m_t^2}{m_W^2}\right)\right]\frac{\alpha(m_W)}{4\pi}\\ \mathcal{Q}_{8(10)} =&(\overline{s}_i b^j)_{V-A} \sum_q \frac{3}{2} Q_q (\overline{q}_j q^i)_{V\pm A}\,,\quad C_{8(10)} = 0\end{aligned} \tag{77}$$

- **Electromagnetic and chromo-magnetic dipole operators:**

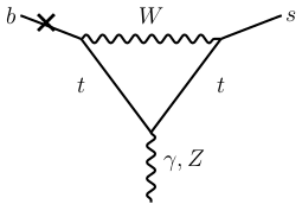


Fig. 11: Topology for electro- and chromo-magnetic dipoles. The cross means mass insertion.

$$\begin{aligned}\mathcal{Q}_{7\gamma} =& -\frac{e}{8\pi^2} m_b\, \overline{s_{Li}} \sigma^{\mu\nu} b_R^i F_{\mu\nu}\,,\quad C_{7\gamma} = -\frac{1}{3} + \mathcal{O}\left(\frac{m_W^2}{m_t^2}\right)\\ \mathcal{Q}_{8g} =& -\frac{g}{8\pi^2} m_b\, \overline{s_{Li}} \sigma^{\mu\nu} (T^a)^i{}_j b_R^j G^a_{\mu\nu}\,,\quad C_{8g} = -\frac{1}{8} + \mathcal{O}\left(\frac{m_W^2}{m_t^2}\right)\end{aligned} \tag{78}$$

- **ΔS = 2 and ΔB = 2 operators**

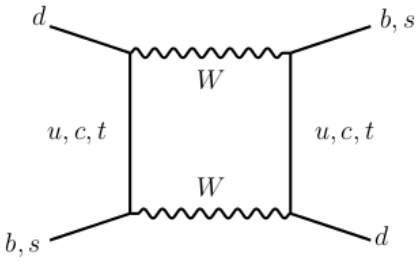


Fig. 12: Box topology

$$\mathcal{Q}(\Delta S = 2) = (\overline{s}_i d^i)_{V_A} (\overline{s}_j d^j)_{V-A}\,,\quad \mathcal{Q}(\Delta B = 2) = (\overline{b}_i d^i)_{V_A} (\overline{b}_j d^j)_{V-A} \tag{79}$$

- **Semileptonic operators:**

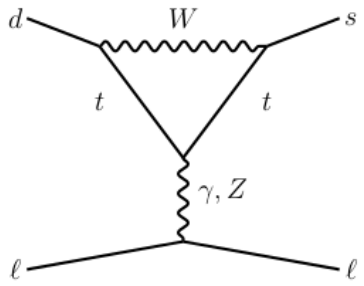


Fig. 13: Semileptonic penguin topology

$$\begin{aligned} \mathcal{Q}_{7V,A} =& (\bar{s}_i d^i)_{V-A}(\bar{e}e)_{V,A}\,, \quad \mathcal{Q}_{9V,10A} = (\bar{s}_i b^i)_{V-A}(\bar{\mu}\mu)_{V,A}\,, \\ \mathcal{Q}_{\bar{\nu}\nu} =& (\bar{s}_i d^i)_{V-A}(\bar{\nu}\nu)_{V-A}\,, \quad \mathcal{Q}_{\bar{\mu}\mu} = (\bar{s}_i d^i)_{V-A}(\bar{\mu}\mu)_{V-A} \end{aligned} \tag{80}$$

With the list of $d = 6$ operators we are able to describe several SM flavour changing processes. For example, the relevant interactions to the parton process $b \to s + \bar{q}q$ can be parametrized though the Hamiltonian

$$\mathcal{H}_{\text{SM}}^{b\to s+q\bar{q}} = -\frac{G_F}{\sqrt{2}}\left[\sum_{p=u,c} V_{pb}^* V_{ps} \sum_{i=1,2} C_i(\mu)\mathcal{Q}_i^p + V_{tb}^* V_{ts} \sum_{i=3,\cdots,10} C_i(\mu) Q_i\right]\,. \tag{81}$$

If we are also interested in $b \to s$ transitions with a photon or a lepton pair in the final state, additional dimension-six operators must be included. We then get,

$$\mathcal{H}_{\text{SM}}^{b\to s+\gamma(l\bar{l})} = \mathcal{H}_{SM}^{b\to s+q\bar{q}} - \frac{G_F}{\sqrt{2}} V_{tb}^* V_{ts}\left[C_{7\gamma}(\mu)Q_{7\gamma} + C_{8g}(\mu)Q_{8g} + C_{9V}(\mu)Q_{9V} + C_{10A}(\mu)Q_{10A}\right]\,. \tag{82}$$

3.3 Effective theories for heavy flavours: a brief introduction

What is there to integrate out, when there are no heavy particles? The answer to this question is in looking for different scales, e.g. in $B-$physics $m_b \gg \Lambda_{\text{QCD}}$. Then we can use the effective theory approach and integrate out all short-distance fluctuations associated with scales $\gg \Lambda_{\text{QCD}}$. In this scenario physics at the m_b scale are short-distance effects, while heavy quark related hadronic physics governed at confinement scale Λ_{QCD} reflect long-distance effects. The separation of the short-distance and long-distance effects associated with these two scales is vital for any quantitative description in heavy-quark physics.

The prime example of this separation is on heavy quark effective field theory (HQET) [19]. What is the physical picture behind HQET?

- Scale hierarchy $m_b \gg \Lambda_{\text{QCD}}$, $\alpha_2(m_B)$ is perturbative (asymptotic freedom)
- Heavy quark - heavy quark system is perturbative
- Heavy-light bound states are not perturbative
- Characterized by a small Compton wavelength; $\lambda_Q \sim 1/m_Q \ll 1/\Lambda_{QCD} \sim R_{had}$(typical hadronic size)

These requirements simplify the physics of hadrons made up of a heavy quark. In mesons composed of a heavy quark, Q, and a light antiquark, $\bar{q}$ (and gluons and $q\bar{q}$ pairs), the heavy quark acts as a static

color source with fixed four-velocity, v_μ , and the wave function of the light degrees of freedom becomes insensitive the mass (flavour) of the heavy quark. Since the magnetic moment of a heavy quark scales like $\mu_Q \sim 1/m_Q$, its spin also decouples. This results in

> **$\mathbf{SU(2n_Q)}$ spin-flavour symmetry:** *In heavy-quark limit ($m_Q \to \infty$), configuration of light degrees of freedom is independent of the spin and flavour of the heavy quark.*

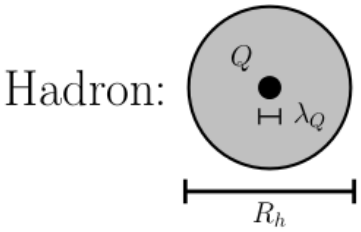


Fig. 14: Pictorial representation of the Hadron. The black central dot represents the heavy quark and the gay are the light degrees of freedom. R_h is the size of the hadron while λ_Q the Compton wave length of the heavy quark.

In the effective description that we are looking there are some other important aspects:

- Heavy quarks carries almost all momentum;
- The momentum exchange between heavy quark and light degrees of freedom is predominantly soft (soft gluon exchange):

$$\Delta P_Q = -\Delta P_{light} = \mathcal{O}(\Lambda_{\rm QCD}) \quad \Rightarrow \quad \Delta v_Q = \mathcal{O}(\Lambda_{\rm QCD}/m_Q)\,; \tag{83}$$

- Heavy-quark velocity becomes a conserved quantum number in $m_Q \to \infty$ limit. This is known as the Georgi "velocity superselection rule";
- Spin doublets such as (B, B^*) should be degenerate in the heavy quark limit: $m_{B^*} - m_B = 46\,\text{MeV} \ll \Lambda_{\rm QCD}$;
- Away from the heavy-quark limit, $1/m_Q$ corrections are expected: $m_{B^*} - m_B = (c_1 - c_0)\lambda_2/m_b + \mathcal{O}(1/m_b^2)$;
- The approach gives a prediction for $(m_{B^*} - m_B)/(m_{D^*} - m_D) \simeq m_c/m_b \simeq 1/3$; Not far from the experimental value of 0.32.

We can now construct an effective theory that makes the effects of the heavy-quark symmetry explicit, i.e. the HQET. The heavy quark Q in the interactions with soft partons (ligh quark q and gluon g) is almost on-shell, such that we can expand the momentum as

$$p_Q^\mu = \underbrace{m_Q v^\mu}_{\substack{\text{hadron}\\ \text{rest frame}\\ v^\mu = (1,0,0,0)}} + \underbrace{k^\mu}_{\substack{\text{residual off-shell}\\ \text{momentum}\\ |k| = \mathcal{O}(\Lambda_{QCD})}} \tag{84}$$

Expanding the heavy quark propagator we get

$$\frac{i}{\not p - m_Q} = \frac{i(\not p + m_Q)}{p^2 - m_Q^2} = \frac{i(m_Q \not v + \not k + m_Q)}{2m_Q v.k + k^2} = \frac{i}{v.k}\frac{1+\not v}{2} + \cdots . \tag{85}$$

We can see that in this expansion the propagator is no longer dependent on the mass of the heavy quark, a clear manifestation of the heavy quark flavour symmetry. To derive the effective Lagrangian is convenient

to decompose the Dirac spinor components into 'upper' (large) and 'lower' (small) pieces

$$Q(x) = e^{-im_Q c.x}[\underbrace{h_v(x) + H_v(x)}_{\text{carry the residual } k}], \quad \text{with} \quad \left[\begin{array}{l} h_v(x) = e^{im_Q v.x} P_+ Q(x) \\ \\ H_v(x) = e^{im_Q v.x} P_- Q(x) \end{array} \right. \tag{86}$$

and $P_\pm = (1 \pm \not{v})/2$ are projector operators. In the rest frame of the heavy quark $P_+ = (1+\gamma^0)/2$ project onto the heavy quark components. An useful identity of these projectors is

$$P_+ \gamma^\mu P_+ = P_+ v^\mu P_+ = v^\mu P_+ \,. \tag{87}$$

Note that $h_v(x)$ and $H_v(x)$ are eigenstates of the velocity operator, i.e. $\not{v} h_v(x) = h_v(x)$ and $\not{v} H_v(x) = -H_v(x)$. In terms of these fields the QCD Lagrangian can now be written as

$$\begin{aligned} \mathcal{L}_{\rm Q} =& \overline{Q}(i\not{D} - m_Q)Q \\ =& \overline{h}_v i\not{D} h_v + \overline{H}_v(i\not{D} - 2m_Q)H_v + \overline{h}_v i\not{D} H_v + \overline{H}_v i\not{D} h_v \\ =& \overline{h}_v iv.D h_v + \overline{H}_v(-iv.D - 2m_Q)H_v + \overline{h}_v i\not{D}_\perp H_v + \overline{H}_v i\not{D}_\perp h_v \end{aligned} \tag{88}$$

where we defined $i\vec{D}^\mu_\perp = iD^\mu - v^\mu iv.D$, orthogonal to the heavy-quark velocity $v.D_\perp = 0$. In the rest frame, $D^\mu_\perp = (0, \vec{D})$ contains the spatial components of the covariant derivative. We see from the Lagrangian above that the component $h_v(x)$ is a massless mode describing a quantum fluctuation around mass-shell, while $H_v(x)$ is a massive mode with mass $2m_Q$ describing a hard quantum fluctuation. This heavy component can be integrated out by using the classical equation of motion

$$H_v = \frac{1}{2m_Q + iv.D} i\not{D}_\perp h_v = \frac{1}{2m_Q} \sum_{n=0}^{\infty} \underbrace{\left(-\frac{iv.D}{2m_Q}\right)^n}_{\substack{\text{small} \\ k \ll m_Q}} i\not{D}_\perp h_v \longrightarrow H_v \simeq \left(\frac{D}{m_Q}\right) h_v \sim \left(\frac{\Lambda_{QCD}}{m_Q}\right) h_v \,. \tag{89}$$

The effective Lagrangian can then be written as

$$\begin{aligned} \mathcal{L}_{\rm HQET} =& \quad \overline{h}_v iv.D_s h_v + \overline{h}_v i\not{D}_\perp \frac{1}{2m_Q + iv.D} i\not{D}_\perp h_v \qquad \longrightarrow \quad \textbf{non-local} \\ =& \quad \overline{h}_v iv.D_s h_v + \frac{1}{2m_Q} \textstyle\sum_{n=0}^{\infty} \overline{h}_v i\not{D}_\perp \left(-\frac{iv.D}{2m_Q}\right)^n i\not{D}_\perp h_v \quad \longrightarrow \quad \textbf{local} \end{aligned} \tag{90}$$

Therefore at leading only $h_v(x)$ contributes, and the effects of $H_v(x)$ are suppressed by powers of $\Lambda_{\rm QCD}/m_Q$, i.e.

$$\mathcal{L}_{\rm HQET} = \overline{h}_v iv.D_s h_v + \mathcal{O}(1/m_Q)\,, \quad \text{with} \quad iD^\mu_s = i\partial^\mu + \underbrace{g_s G^\mu_s}_{\text{soft gluons}} \,. \tag{91}$$

It is straightforward to extend the above result for higher order of power corrections. At the next to leading order we get

$$\mathcal{L}_{\rm HQET} = \underbrace{\overline{h}_v iv.D_s h_s}_{\substack{SU(2n_Q) \\ \text{spin-flavour} \\ \text{symmetry}}} + \frac{1}{2m_Q}[\underbrace{\overline{h}_v (i\vec{D}_s)^2 h_v}_{\substack{-\overline{h}_v(i\vec{D}_s)^2 h_v \\ \uparrow \\ \text{kinetic-energy} \\ \text{operator}}} + \underbrace{C_{\rm mag}(\mu)\frac{g_s}{2}\overline{h}_v \sigma_{\mu\nu} G^{\mu\nu}_s h_v}_{\substack{-4\overline{h}_v \vec{S}.\vec{B}_c h_v \\ \uparrow \\ \text{chromo-magnetic} \\ \text{from pert. theo.}}}] + \cdots \,, \tag{92}$$

where we have make use of the identity

$$P_+ i\not{D}_\perp i\not{D}_\perp P_+ = P_+ \left[(iD_\perp)^2 + \frac{g_s}{2}\sigma_{\mu\nu}G^{\mu\nu}\right] P_+ \tag{93}$$

and $i[D^\mu, D^\nu] = g_s G^{\mu\nu}$ is the gluon fields-strength tensor. Here $\vec{S}$ is the spin operator and $B_c^i = -1/2\epsilon^{ijk}G^{jk}$ are the components of the colour-magnetic field. The Wilson coefficient is computed through RGE-improved perturbation theory [33]. The leading term is $SU(2n_Q)$ spin-flavour invariant, i.e. no reference to the heavy-quark mass (flavour symmetry) and invariant under the spin rotations $h_v \to (1 + i/2\vec{\epsilon}.\vec{\sigma})h_v$. The flavour symmetry is broken by the operators arising at order $1/m_Q$ and higher. Note, however, that at this order the kinetic term conserves the spin symmetry, while the chromo-magnetic operator breaks the both flavour and spin symmetry. Figure 15 shows the changes in the Feynman rules in the new formalism.

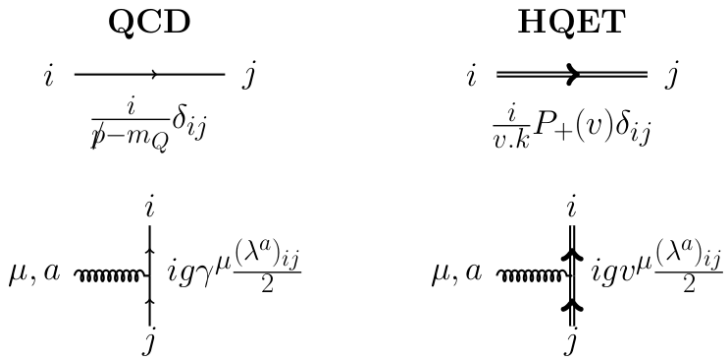


Fig. 15: Feynman rules QCD vs. HQET

Up to now we have integrated out small components in the heavy-quark fields and obtained an effective local Lagrangian that describes the long-distance physics in the full theory. The way heavy-quarks participate in the strong interaction is through their couplings to gluons. These can be soft (virtual momentum small, of the order of the confining scale) or hard (virtual momentum large, of the order of the heavy quark mass). In the approach used above we have integrated out the hard gluons as they, contrarily to the soft ones, break the heavy-quark symmetries. However, hard gluons are important once we decide to add short-distance effects. Their effects lead to a renormalization of the coefficients of the operators in the HQET Lagrangian, which are calculable in perturbation theory. There is no renormalization at leading order. Nor renormalization of the kinetic operator due to Lorentz invariance ("reparametrization invariance"). However, the chromo-magnetic interaction will be affected.

Heavy-quark symmetry is particularly predictive for exclusive semi-leptonic B decays such as $B \to D^{(*)}\ell\bar{\nu}$. It allow us to extract the CKM matrix elements $|V_{cb}|$ and $|V_{ub}|$ with controlled theoretical uncertainties, through the correlations shown in Fig. 16 .

A clever use of heavy-quark symmetries allows us to calculate the decay rate at the special kinematic point of maximum momentum transfer to the leptons ($v = v'$), i.e. "zero recoil" point. How can we deal with confinement effects in this hadronic process? We can consider elastic scattering of a B meson, $\bar{B}(v) \to \bar{B}(v')$, induced by the vector current $J^\mu = \bar{b}\gamma^\mu b$. The heavy quark acts as a static source of color, and the light quarks orbit around it before the action of the vector current. On average, the b quark and the B meson have the same velocity. The action of the current is to replace instantaneously (at $t = t_0$) the color source by one moving at speed v'. Nothing happens if $v = v'$, i.e. the final state remains a B meson with probability 1 (case (a) in Fig. 17). However, for $v \neq v'$, the probability for an elastic transition is less than 1. The light constituents find them selfs interacting with moving source. Soft gluons will have to be exchanged in order to rearrange them and form a B meson moving at a different speed, leading to a form factor suppression. In the Heavy-quark mass limit, i.e. $m_b \to \infty$, the process is

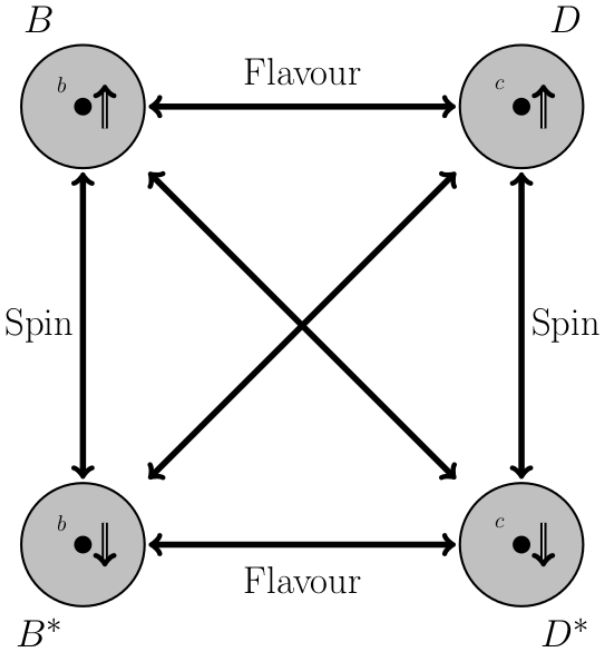


Fig. 16: Spin-flavour symmetry between B- and D-system

described by a dimensionless probability function $\xi(v.v')$ called the Isgur-Wise function. The hadronic matrix elements describing the scattering process is then

$$\frac{1}{m_B}\langle \bar{B}(v')|\bar{b}_{v'}\gamma^\mu b_v|\bar{B}(v)\rangle = \xi(v.v')(v+v')^\mu\,, \quad \text{with} \quad \xi(v.v') \leq 1\,,\, \xi(1) = 1\,. \tag{94}$$

The $1/m_B$ factor on the left-hand side of the equation compensates the normalization of the meson state, i.e. $\langle \bar{B}(p')|\bar{B}(p)\rangle = 2m_B v^0(2\pi)^3\delta(\vec{p}-\vec{p}')$. We can then use the flavour symmetry to replace $b-$ by $c-$quark in the final state, thereby obtaining a $B \to D$ transition. This transforms the scattering process into a weak decay process.

Nothing will happen to the matrix element since in the heavy-quark limit the Lagrangian is invariant under the $b_{v'} \to c_{v'}$ replacement (case (b) in Fig. 17), i.e.

$$\frac{1}{\sqrt{m_B m_D}}\langle \bar{D}(v')|\bar{c}_{v'}\gamma^\mu b_v|\bar{B}(v)\rangle = \xi(v.v')(v+v')^\mu\,. \tag{95}$$

This is a very interesting prediction of the heavy-quark symmetry. Since in general the matrix element of a flavour-changing current between two pseudo-scalar mesons is given by

$$\langle \bar{D}(v')|\bar{c}_{v'}\gamma^\mu b_v|\bar{B}(v)\rangle = f_+(q^2)(p+p')^\mu - f_-(q^2)(p-p')^\mu\,, \tag{96}$$

with $f_\pm(q^2)$ the form factors and $q = p - p'$. The heavy-quark symmetry relates the two a priori independent form factors to one and the same function, i.e. the Isgur-Wise function ($f_\pm(q^2) \propto \xi(v.v')$).

Next, we can use the spin symmetry to flip the spin of $c-$quark in final state, thereby obtaining a $B \to D^*$ transition (case (c) in Fig. 17). The current gets transformed to

$$\langle D^*(v',\epsilon)|\bar{c}_{v'}\gamma^\mu(1-\gamma_5)b_v|B(v)\rangle = \langle D^*(v',\epsilon)|\bar{c}_{v'}\gamma^\mu b_v|B(v)\rangle - \langle D^*(v',\epsilon)|\bar{c}_{v'}\gamma^\mu\gamma_5 b_v|B(v)\rangle \tag{97}$$

with

$$\begin{aligned}
\frac{1}{\sqrt{m_B m_{D^*}}}\langle D^*(v',\epsilon)|\bar{c}_{v'}\gamma^\mu b_v|B(v)\rangle &= i\epsilon^{\mu\nu\alpha\beta}\epsilon^*_\nu v'_\alpha v_\beta \xi(v.v') \\
\frac{1}{\sqrt{m_B m_{D^*}}}\langle D^*(v',\epsilon)|\bar{c}_{v'}\gamma^\mu\gamma_5 b_v|B(v)\rangle &= [\epsilon^{*\mu}(v.v'+1) - v'^\mu \epsilon^*.v]\xi(v.v')
\end{aligned} \tag{98}$$

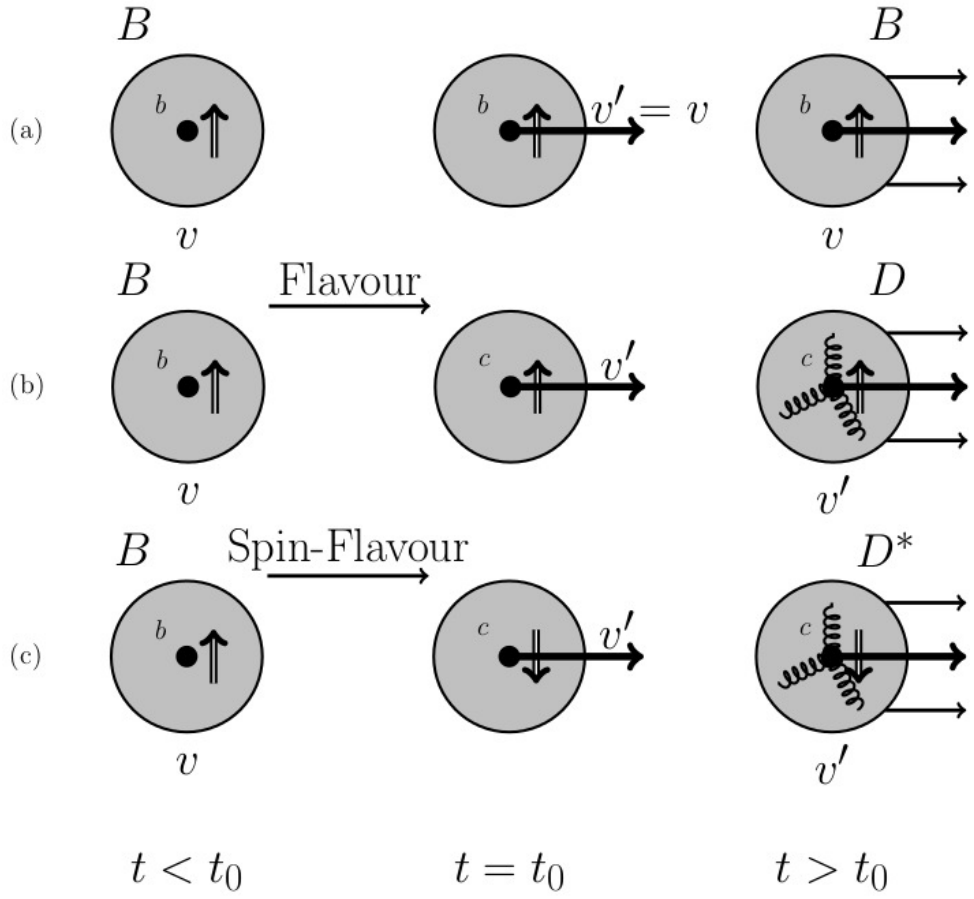


Fig. 17: Evolution with time of the hadron for the different scenarios where spin-flavour symmetry is applied.

where ϵ denotes the polarization of the D^* meson. The general Lorentz-invariant matrix elements of these hadron currents are given by

$$\begin{aligned}\langle D^*(v',\epsilon)|\bar{c}_{v'}\gamma^\mu b_v|B(v)\rangle =& \frac{2i}{m_B+m_{D^*}}\epsilon_{\mu\nu\alpha\beta}\epsilon^{*\nu}p'^\alpha p^\beta V(q^2)\\ \langle D^*(v',\epsilon)|\bar{c}_{v'}\gamma^\mu\gamma_5 b_v|B(v)\rangle =& (m_B+m_{D^*})\epsilon^*_\mu A_1(q^2) - \frac{\epsilon^*.p}{m_B+m_{D^*}}(p+p')_\mu A_2(q^2)\\ & - 2m_{D^*}\frac{\epsilon^*.q}{q^2}q_\mu A_3(q^2) + 2m_{D^*}\frac{\epsilon^*.q}{q^2}q_\mu A_0(q^2)\end{aligned} \tag{99}$$

with

$$A_3(q^2) = \frac{m_B+m_D}{2m_{D^*}}A_1(q^2) - \frac{m_B-m_{D^*}}{2m_{D^*}}A_2(q^2)\,. \tag{100}$$

In general, these exclusive semileptonic decays processes can be described by six a priori independent

hadronic form factors

$$\left[\begin{array}{l} \textbf{For } \mathbf{0^- \to 0^-} \textbf{ transition: } I \to F\ell\nu_\ell \\ \\ \langle F(v')|V_\mu^{cb}|I(v)\rangle = \sqrt{m_I m_F}\,[\xi_+(v.v')(v+v')_\mu + \xi_-(v.v')(v-v')_\mu] \end{array}\right.$$

$$\left[\begin{array}{l} \textbf{For } \mathbf{0^- \to 1^-} \textbf{ transition: } I \to F^*\ell\nu_\ell \\ \\ \langle F^*(v')|V_\mu^{cb}|I(v)\rangle = i\sqrt{m_I m_{F^*}}\xi_V(v.v')\epsilon_{\mu\nu\alpha\beta}\epsilon^{\mu\nu}v'^\alpha v^\beta \\ \langle F^*(v')|A_\mu^{cb}|I(v)\rangle = \sqrt{m_I m_{F^*}}[\xi_{A_1}(v.v')(v.v'+1)\epsilon_\mu^* - \xi_{A_2}(v.v')\epsilon^*.v v_\mu \\ \qquad\qquad -\xi_{A_3}(v.v')\epsilon^*.v v'_\mu]\,. \end{array}\right. \tag{101}$$

with V_μ and A_μ the vector- and axial-currents, respectively. The heavy-quark limit imposes the relations:

$$\xi_+(v.v') = \xi_V(v.v') = \xi_{A_1}(v.v') = \xi_{A_3}(v.v') = \xi(v.v') \quad \text{and} \quad \xi_-(v.v') = \xi_{A_2}(v.v') = 0\,. \tag{102}$$

These relations are model independent and are a consequence of QCD in the limit $m_b, m_c \gg \Lambda_{\rm QCD}$. For the processes described below the form factor correlations read

$$\begin{aligned} \xi(v.v') &= \frac{2\sqrt{m_B m_D}}{m_B \pm m_D} f_\pm(q^2) = \frac{2\sqrt{m_B m_{D^*}}}{m_B + m_{D^*}} V(q^2) = \frac{2\sqrt{m_B m_{D^*}}}{m_B + m_{D^*}} A_0(q^2) \\ &= \frac{2\sqrt{m_B m_{D^*}}}{m_B + m_{D^*}} A_2(q^2) = \frac{2\sqrt{m_B m_{D^*}}}{m_B + m_{D^*}} \left[1 - \frac{q^2}{(m_B + m_{D^*})^2}\right]^{-1} A_1(q^2)\,, \end{aligned} \tag{103}$$

with $q^2 = m_B^2 + m_{D^*}^2 - 2m_B m_{D^*} v.v'$. These from factors play an important role in describing semileptonic decays as $\bar{B} \to D^{(*)}\ell\nu$. In terms of the recoil variable $\omega = v.v'$, the differential decay rate in the heavy quark limit for these processes is given by

$$\frac{d\Gamma(B \to D^{(*)}\ell\bar{\nu})}{d\omega} = \frac{G_F^2 \eta_{ew}^2}{48\pi^3}|V_{cb}|^2 \times F \times \left\{\begin{array}{ll} (\omega^2-1)^{1/2}\mathcal{F}_*^2(\omega)\,, & \text{for } B \to D^* \\ \\ (\omega^2-1)^{3/2}\mathcal{F}^2(\omega)\,, & \text{for } B \to D \end{array}\right. , \tag{104}$$

with $\eta_{ew} \simeq 1$ a parameter accounting for the electroweak corrections to the four-fermion operator mediating the decay and

$$F = \left\{\begin{array}{ll} m_B^5 r^3(1-r)^2(\omega+1)^2\left(1 + \dfrac{4\omega}{\omega+1}\dfrac{1-2r\omega+r^2}{(1-r)^2}\right),\ r = \dfrac{m_{D^*}}{m_B} & \text{for } D^* \\ \\ (m_B + m_D)^2 m_D^3 & \text{for } D \end{array}\right. \tag{105}$$

Both $\mathcal{F}(\omega)$ and $\mathcal{F}_*(\omega)$ are equal in the heavy-quark mass limit and are normalized such that $\mathcal{F}_{(*)}(1) = 1$, allowing a model independent extraction of $|V_{cb}|$. The above differential decay rate expressions receive symmetry-breaking corrections, since the mass of the heavy quark is not infinitely large:

- Corrections of order $\mathcal{O}(\alpha_s^n(m_Q))$ (hard gluons) can be calculated perturbatively;
- Power corrections of order $\mathcal{O}((\Lambda_{\rm QCD}/m_Q)^n)$ are non-perturbative and more difficult to control.

These corrections have been estimated and schematically give

$$\begin{aligned}\mathcal{F}_*(1) \simeq & 1 + \underbrace{c_A(\alpha_s)}_{\text{Perturbative}} + \overbrace{0 \times \frac{\Lambda_{\rm QCD}}{m_Q}}^{\substack{\text{Luke}\\ \text{Theorem}}} + \overbrace{\text{cons} \times \frac{\Lambda^2_{\rm QCD}}{m^2_Q}}^{\substack{\text{lattice/}\\ \text{models}}} + \cdots \\ \mathcal{F}(1) \simeq & 1 + \overbrace{c_V(\alpha_s)}^{\text{Perturbative}} + \underbrace{\text{const} \times \frac{\Lambda_{\rm QCD}}{m_Q}}_{\substack{\text{lattice/}\\ \text{models}}} + \cdots\end{aligned} \tag{106}$$

The absence of the $\mathcal{O}(\Lambda_{\rm QCD}/m_Q)$ term for $B \to D^* \ell \bar{\nu}_\ell$ at the zero-recoil limit, i.e. $\omega = 1$, is a consequence of the Luke theorem:

> *The matrix elements describing the leading $1/m_Q$ corrections to weak decay amplitudes vanish at zero recoil, to all order in perturbation theory.*

The reason why in the semi-leptonic decay $B \to D \ell \bar{\nu}_\ell$ this is no longer true is more subtle and can be found in [34]. Therefore, from the value of $\mathcal{F}_*(1)$ the value of $|V_{cb}|$ is estimated to be

$$\begin{aligned} |V_{cb}| &= (39.48 \pm 0.5_{exp} \pm 0.74_{theo}) \times 10^{-3} \quad \text{from lattice QCD}\,, \\ |V_{cb}| &= (41.4 \pm 0.5_{exp} \pm 1.0_{theo}) \times 10^{-3} \quad \text{from QCD sum rules}\,, \end{aligned} \tag{107}$$

showing the power of HQET in describing non-pertubative systems.

4 Some aspect of CP violation

4.1 CP violation in the Universe

One of the currents issues related with flavour physics and CP violation is the Baryon asymmetry of the Universe. Our understanding of the Universe is based on the Standard Cosmological Model, where the Universe expanded from a primordial hot and dense initial state at some finite time in the past (the so-called Big Bang) and is then followed by a period of inflationary expansion that ensured the curvature to become approximately zero [35]. After this inflationary epoch, the Universe continued to expand but at a low rate. The rate of expansion is determined by the component of energy density that dominates the total energy density; at the present time this is the so-called dark energy component, which causes the expansion to accelerate due to its negative pressure.

In our surroundings the objects are mostly made of matter, e.g. planets, stars, etc.. The present value of the baryon asymmetry of the Universe inferred from WMAP seven-year data combined with baryon acoustic oscillations is [36]

$$\eta_B \equiv \frac{n_B - n_{\bar{B}}}{n_\gamma} = (6.19 \pm 0.14) \times 10^{-10}\,, \tag{108}$$

where n_B, $n_{\bar{B}}$ and n_γ are the number density of baryons, antibaryons and photons at present time, respectively. The smallness of this quantity poses a challenge to both particle physics and cosmology. If we take inflation for granted, then in the early Universe any primordial cosmological asymmetry would be erased during the inflationary period. This is one argument that strongly suggests this asymmetry to be dynamically generated, instead of being an initial accidental state. Sakharov realized the need of three ingredients in order to create a baryon asymmetry from an initial state with baryon number equal to zero [37]. The three conditions can be stated as follows:

i) Baryon number violation;
ii) C and CP violation;
iii) Departure from thermal equilibrium.

The first condition is rather obvious. If there is no B violation, the baryon number is conserved in all interactions and, therefore, commutes with the Hamiltonian at any time, i.e.

$$[B,\mathcal{H}]=0 \quad \Rightarrow \quad B(t)=\int_0^t [B,\mathcal{H}]dt'=0\,. \tag{109}$$

The second condition is a little more delicate. Let us start by writing the baryon number operator

$$\hat{B}=\frac{1}{3}\sum_i \int d^3x \;:\psi_i^\dagger(\vec{x},t)\psi_i(\vec{x},t):, \tag{110}$$

where $\psi_i(\vec{x},t)$ denotes the quark field of flavour i and :: denote the normal ordering. The C, P and T transformations of these fields are given in Tables 1–2. Thus, the fermionic number satisfies the following transformations

$$\begin{aligned}
\mathcal{P}:\psi_i^\dagger(\vec{x},t)\psi_i(\vec{x},t):\mathcal{P}^{-1}&=:\psi_i^\dagger(-\vec{x},t)\psi_i(-\vec{x},t):,\\
\mathcal{C}:\psi_i^\dagger(\vec{x},t)\psi_i(\vec{x},t):\mathcal{C}^{-1}&=-:\psi_i^\dagger(\vec{x},t)\psi_i(-\vec{x},t):,\\
\mathcal{T}:\psi_i^\dagger(\vec{x},t)\psi_i(\vec{x},t):\mathcal{T}^{-1}&=:\psi_i^\dagger(\vec{x},-t)\psi_i(\vec{x},-t):\;.
\end{aligned} \tag{111}$$

We can, therefore, find how the baryon number operator transforms under these operators. One gets

$$\mathcal{C}\hat{B}\mathcal{C}^{-1}=-\hat{B}\,,\quad (\mathcal{CP})\hat{B}(\mathcal{CP})^{-1}=-\hat{B}\,,\quad (\mathcal{CPT})\hat{B}(\mathcal{CPT})^{-1}=-\hat{B}\,. \tag{112}$$

Now, if C is conserved, then $[\mathcal{C},\mathcal{H}]=0$ and the expectation value of the baryon number is given by

$$\begin{aligned}
\left\langle\hat{B}(t)\right\rangle&=\left\langle e^{i\mathcal{H}t}\hat{B}(0)e^{-i\mathcal{H}t}\right\rangle=\left\langle\mathcal{C}^{-1}\mathcal{C}e^{i\mathcal{H}t}\hat{B}(0)e^{-i\mathcal{H}t}\right\rangle=\left\langle e^{i\mathcal{H}t}\mathcal{C}\hat{B}(0)\mathcal{C}^{-1}e^{-i\mathcal{H}t}\right\rangle\\
&=-\left\langle e^{i\mathcal{H}t}\hat{B}(0)e^{-i\mathcal{H}t}\right\rangle=-\left\langle\hat{B}(t)\right\rangle\,.
\end{aligned} \tag{113}$$

We see that the expectation value $\left\langle\hat{B}(t)\right\rangle$ is only different from zero if C is not a symmetry of the Hamiltonian. The same is true for CP.

The last condition can be understood as follows. In thermal equilibrium, the thermal average are weighted by the density operator $\rho=e^{-\beta\mathcal{H}}$, with $\beta=1/T$. Assuming CPT invariance of the Hamiltonian we get

$$\begin{aligned}
\left\langle\hat{B}(t)\right\rangle_T&=\text{Tr}\left[e^{\beta\mathcal{H}}\hat{B}\right]=\text{Tr}\left[(\mathcal{CPT})^{-1}(\mathcal{CPT})e^{\beta\mathcal{H}}\hat{B}\right]=\text{Tr}\left[e^{\beta\mathcal{H}}(\mathcal{CPT})\hat{B}(\mathcal{CPT})^{-1}\right]\\
&=-\left\langle\hat{B}(t)\right\rangle_T\,.
\end{aligned} \tag{114}$$

This means that, within a CPT invariant Hamiltonian, the thermal average is zero and no net baryon asymmetry is generated since the inverse processes will destroy the asymmetry generated in the direct decays. Departure from thermal equilibrium is very common in the early Universe when interaction rates cannot keep up with the expansion rate of the Universe.

All three of these condition can be found in the SM, however the amount of CP violation from the CKM mechanisms is to small in order to generate such an asymmetry.

4.2 Weak and strong phases

CP is violated in nature by the weak interactions. The imposition of CP invariance in a transition amplitude is expressed as

$$(\mathcal{CP})\,\widehat{T}\,(\mathcal{CP})^{\dagger} = \widehat{T}\,. \tag{115}$$

In classical physics, the square of the *CP* transformation is identical to the identity transformation, and therefore $(\mathcal{CP})^2$ corresponds to a conserved quantum number. The value of $(\mathcal{CP})^2$ for initial and final states must be identical, and is a purely arbitrary phase. Without loss of generality one can choose $(\mathcal{CP})^2 = 1$. The *CP* transformations read

$$\mathcal{CP}\,|i\rangle = e^{i\xi_i}\,\left|\bar{i}\right\rangle\,, \quad \mathcal{CP}\,\left|\bar{i}\right\rangle = e^{-i\xi_i}\,|i\rangle\,, \tag{116}$$

with ξ_i an arbitrary phase. The CP constraints on the transition amplitudes from an initial state i to the final states f and g are

$$\begin{array}{c}\textbf{Final state}\\ f/\bar{f}\end{array}\left[\begin{array}{l}\langle f|\,\widehat{T}\,|i\rangle = e^{i(\xi_i-\xi_f)}\,\left\langle\overline{f}\right|\widehat{T}\,\left|\bar{i}\right\rangle\\[2ex] \left\langle\overline{f}\right|\widehat{T}\,|i\rangle = e^{i(\xi_i+\xi_f)}\,\langle f|\,\widehat{T}\,\left|\bar{i}\right\rangle\end{array}\right.,\quad \begin{array}{c}\textbf{Final state}\\ g/\bar{g}\end{array}\left[\begin{array}{l}\langle g|\,\widehat{T}\,|i\rangle = e^{i(\xi_i-\xi_g)}\,\langle\overline{g}|\,\widehat{T}\,\left|\bar{i}\right\rangle\\[2ex] \langle\overline{g}|\,\widehat{T}\,|i\rangle = e^{i(\xi_i+\xi_g)}\,\langle g|\,\widehat{T}\,\left|\bar{i}\right\rangle\end{array}\right. . \tag{117}$$

From these transition amplitudes one sees that the modulus of each process is equal to the modulus of the CP conjugated one. Therefore, the CP-violating quantities are

$$\begin{array}{c}\textbf{Final state}\\ f/\bar{f}\end{array}\left[\begin{array}{l}\left|\langle f|\,\widehat{T}\,|i\rangle\right| - \left|\left\langle\overline{f}\right|\widehat{T}\,\left|\bar{i}\right\rangle\right|\\[2ex] \left|\left\langle\overline{f}\right|\widehat{T}\,|i\rangle\right| - \left|\langle f|\,\widehat{T}\,\left|\bar{i}\right\rangle\right|\end{array}\right.,\quad \begin{array}{c}\textbf{Final state}\\ b/\bar{b}\end{array}\left[\begin{array}{l}\left|\langle g|\,\widehat{T}\,|i\rangle\right| - \left|\langle\overline{g}|\,\widehat{T}\,\left|\bar{i}\right\rangle\right|\\[2ex] \left|\langle\overline{g}|\,\widehat{T}\,|i\rangle\right| - \left|\langle g|\,\widehat{T}\,\left|\bar{i}\right\rangle\right|\end{array}\right. \tag{118}$$

$$\begin{array}{ccccc}\downarrow & & & & \downarrow\\ \neq 0 & \Longrightarrow & \textbf{CP violation} & \Longleftarrow & \neq 0\end{array}$$

If we only had one final state, say f, the relevant expressions would be the ones presented in the first line of Eq. (117) and (118). In Eq. (117), we only have two phases for two complex equations and therefore no other quantity beyond the one presented in Eq. (118) would violate CP. The fact that we have two final states, f and g, leads to three arbitrary phases but four complex equations. Since we only have four real CP-violating quantities in Eq. (118), a physical CP condition on the phases of the decay amplitudes must remain. One can find that the quantity

$$\langle f|\,\widehat{T}\,|i\rangle\,\left\langle\overline{f}\right|\widehat{T}\,|i\rangle\,\langle g|\,\widehat{T}\,\left|\bar{i}\right\rangle\,\langle g|\,\widehat{T}\,\left|\bar{i}\right\rangle - \langle g|\,\widehat{T}\,|i\rangle\,\langle\overline{g}|\,\widehat{T}\,|i\rangle\,\langle f|\,\widehat{T}\,\left|\bar{i}\right\rangle\,\langle f|\,\widehat{T}\,\left|\bar{i}\right\rangle \tag{119}$$

must vanish if CP invariance holds.

The presence of complex phases is closely related with CP violation. One simple argument to support this statement is due to CPT invariance. If CPT is conserved then CP violation is the same as T violation. Since T transforms a number into its complex conjugate, the CP violation must be related to the presence of complex numbers. One should stress, however, that the phase of a transition amplitude is arbitrary and non-physical, due to the freedom of phase redefinition of the kets and bras. Only phases which are rephasing invariant can lead to CP violation. These are in general relative phases of transition amplitudes. There are three types of phases that can arise in transitions amplitudes:

- **'weak' or CP-odd phases.**
  The weak phases are defined as the phases that change sign under CP conjugation, and usually originate from complex couplings in the Lagrangian.
- **'strong' or CP-even phases.**
  The strong phases are the ones that remain unchanged under CP conjugation. They may arise from the trace of products of an even number of γ matrices together with γ_5, or final-state-interaction

scatterings from on-shell states. The last one appears when the total amplitude for the decay $i \to f$ includes contributions from $i \to f' \to f$, where the decay $i \to f'$ is through weak interactions and $f' \to f$ through strong or electromagnetic ones. If the intermediate states are on mass shell this creates an absorptive part. These are also typical phases appearing on absorptive parts of loops diagrams in perturbation theory.

- **'spurious' CP-transformation phases.**
The spurious phases are global, purely conventional relative phases between the amplitude of a process and the amplitude for the CP-conjugate process. These phases do not originate in any dynamics, they just come from the assumed CP transformation of the field operators and on the kets and bras they act upon [2].

4.3 Types of CP Violation

- **CP-violation in Decays (direct CP violation)**
This type of CP-violation occurs when a meson P and its CP-conjugate decay at different rates to the same final state (up to CP conjugacy). This can be characterized by the relation

$$\left|\frac{\overline{A}_{\bar{f}}}{A_f}\right| \neq 1\,. \tag{120}$$

In charged meson decays, where mixing is not present, this is the only source of CP violation:

$$a_{f^\pm} = \frac{\Gamma(P^- \to f^-) - \Gamma(P^+ \to f^+)}{\Gamma(P^- \to f^-) + \Gamma(P^+ \to f^+)} = \frac{|\overline{A}_{\bar{f}}/A_f|^2 - 1}{|\overline{A}_{\bar{f}}/A_f|^2 + 1}\,. \tag{121}$$

In order to have CP violation in transition amplitudes from i ($\bar{i}$) to f ($\bar{f}$), the transition amplitudes need to be a sum of two or more interfering amplitudes. The way we can see this is through an explicit example. Consider for instance

$$\langle f|\,T\,|i\rangle = Ae^{i(\delta+\phi)}\,, \quad \left\langle \bar{f}\right|T\left|\bar{i}\right\rangle = Ae^{i(\delta-\phi+\theta)}\,, \tag{122}$$

with A a real positive number, δ a strong phase, ϕ a weak phase and θ a spurious one. It is easy to see that these transition amplitudes satisfy the first equation of Eq. (117) with

$$\xi_i - \xi_f = 2\phi - \theta\,, \tag{123}$$

leading to

$$|\langle f|\,T\,|i\rangle| - \left|\left\langle \bar{f}\right|T\left|\bar{i}\right\rangle\right| = A - A = 0\,. \tag{124}$$

Therefore, no CP violation is generated in such a transition. This is no longer true when there is interference. For that, we consider

$$\begin{aligned}\langle f|\,T\,|i\rangle =& A_1 e^{i(\delta_1+\phi_1)} + A_2 e^{i(\delta_2+\phi_2)}\,,\\ \left\langle \bar{f}\right|T\left|\bar{i}\right\rangle =& A_1 e^{i(\delta_1-\phi_1+\theta_1)} + A_2 e^{i(\delta_2-\phi_2+\theta_2)}\,,\end{aligned} \tag{125}$$

where δ_i, ϕ_i and θ_i are the strong, weak and spurious phases, respectively. Now, it is no longer possible to satisfy Eq. (117). We can evaluate the CP-violating quantity

$$\frac{|\langle f|\,T\,|i\rangle|^2 - \left|\left\langle \bar{f}\right|T\left|\bar{i}\right\rangle\right|^2}{|\langle f|\,T\,|i\rangle|^2 + \left|\left\langle \bar{f}\right|T\left|\bar{i}\right\rangle\right|^2} = \frac{-4A_1A_2\sin(\delta_1-\delta_2)\sin(\phi_1-\phi_2)}{2A_1^2+2A_2^2+4A_1A_2\cos(\delta_1-\delta_2)\cos(\phi_1-\phi_2)}\,. \tag{126}$$

This expression will be used later on (in a different form) and, therefore, it is useful to make a few remarks:

– The existence of both weak and strong phases is crucial for CP violation;
– Only relative phases (weak and strong) are relevant in physical processes;
– The limiting case $|\phi_1 - \phi_2| = |\delta_1 - \delta_2| = \pi/2$ and $A_1 = A_2$ gives the maximum value of the CP asymmetry;

It is possible to have CP violation without strong phases, if we have more than one final state and its CP conjugate. For example, having the transition amplitudes

$$\begin{aligned} \langle f|\,T\,|i\rangle &= A_1 e^{i(\delta_1+\phi_1)}\,, \quad \langle f|\,T\,|\bar{i}\rangle = A_1 e^{i(\delta_1-\phi_1+\theta)}\,, \\ \langle g|\,T\,|i\rangle &= A_2 e^{i(\delta_2+\phi_2)}\,, \quad \langle g|\,T\,|\bar{i}\rangle = A_2 e^{i(\delta_2-\phi_2+\theta)}\,, \end{aligned} \tag{127}$$

with $f = \bar{f}$ and $g = \bar{g}$, we can build the quantity

$$\langle f|\,T\,|\bar{i}\rangle\,\langle g|\,T\,|\bar{i}\rangle - \langle g|\,T\,|i\rangle\,\langle f|\,T\,|i\rangle = 2iA_1A_2e^{i(\delta_1+\delta_2+\theta)}\sin(\phi_1-\phi_2)\,. \tag{128}$$

In this quantity the strong phases are basically irrelevant and CP violation is dictated by the weak phases. However, these two distinct final states must be correlated such that the decay involve both simultaneously, otherwise this can not be an observable. This is actually the case in kaon de decays to $\pi^+\pi^-$ and $\pi^0\pi^0$ (see Sec. 4.4).

- **CP-violation in mixing (indirect CP violation)**
This type of CP violation occurs when degenerated neutral mesons are not the CP eigenstates. This can be characterized by the relation

$$\left|\frac{q}{p}\right| \neq 1\,. \tag{129}$$

This is the only source of CP violation in semileptonic final states such as $P^0 \to l^+X$. In such a scenario the asymmetry can be observed in

$$a_{SL} = \frac{\Gamma(\overline{P^0}_{phys}(t) \to l^+X) - \Gamma(P^0_{phys}(t) \to l^-\bar{X})}{\Gamma(\overline{P^0}_{phys}(t) \to l^+X) + \Gamma(P^0_{phys}(t) \to l^-\bar{X})} = \frac{1-|q/p|^2}{1+|q/p|^2}\,. \tag{130}$$

The meson $P^0_{phys}(t)$ represents the time evolved state. As we shall see in Sec. 4.4,

$$a_{SL} = \text{Im}\left(\frac{\Gamma_{12}}{M_{12}}\right). \tag{131}$$

This means that in our model we just need to know M_{12} and Γ_{12}, in order to compute the CP violating observable. However, in general Γ_{12} is plagued with large hadronic uncertainties, making this computation more cumbersome.

- **CP-violation in interference decays**
This type of CP violation only occurs in decays where the final state f is common for both P^0 and $\overline{P^0}$. This can be characterized by the relation

$$\text{Im}\,\lambda_f \neq 0\,, \tag{132}$$

where $\lambda_f = (q/p)(A(\overline{P^0} \to f_{CP})/A(P^0 \to f_{CP}))$. One example is where this asymmetry can be observed is in decays involving $\mathcal{CP}$ eigenstates with ± 1 eigenvalues. Then we have the CP violating observable

$$a_{f_{CP}}(t) = \frac{\Gamma(\overline{P^0} \to f_{CP}) - \Gamma(P^0 \to f_{CP})}{\Gamma(\overline{P^0} \to f_{CP}) + \Gamma(P^0 \to f_{CP})}\,. \tag{133}$$

In the B-system this leads to

$$a_{f_{CP}}(t) = -\frac{1-|\lambda_{f_{CP}}|^2}{1+|\lambda_{f_{CP}}|^2}\cos(\Delta m_B t) + \frac{2\mathrm{Im}\,\lambda_{f_{CP}}}{1+|\lambda_{f_{CP}}|^2}\sin(\Delta m_B t) \tag{134}$$

The fist term on th l.h.s. corresponds to $\mathcal{CP}$ violation through mixing, while the last term is due to interference. In decays with $|\lambda_{CP}| = 1$ only the interference effect survives

$$a_{f_{CP}}(t) = \mathrm{Im}\,\lambda_{f_{CP}}\sin(\Delta m_B t)\,. \tag{135}$$

We know Δm_B so we can measure $\mathrm{Im}\,\lambda_{f_{CP}}$. This quantity is the phase between mixing and decay amplitudes. To a good approximation $|A(\overline{P^0} \to f_{CP})| = |A(P^0 \to f_{CP})|$ and since in the standard parametrization $q/p = e^{i2\beta}$, we have to a good approximation

$$\mathrm{Im}\,\lambda_{f_{CP}} = \mathrm{Im}\left[\frac{q}{p}\frac{A(\overline{P^0} \to f_{CP})}{A(P^0 \to f_{CP})}\right] \simeq \sin 2\beta\,. \tag{136}$$

4.4 Neutral Meson Mixing: General description

In this section we shall follow closely the the discussions in [9]. We are interested in describing how CP violation arises from the mixing of a neutral meson P_0 with its antiparticle $\overline{P^0}$. Consider the simplest scenario where the two states $|P^0\rangle$ and $|\overline{P^0}\rangle$ that are degenerated can neither decay or transform into each other. In such a system an arbitrary state can then be represented as

$$|\psi(t)\rangle = a(t)|P^0\rangle + b(t)|\overline{P^0}\rangle \tag{137}$$

and evolve through the Schrodinger equation with diagonal Hamiltonian. This scenario is exactly what happens in the neutral meson system when only QCD interactions are active. Turning on the electroweak interactions will induce, even if small, off-diagonal Hamiltonian entries mixing both states leading to the breaking of the degeneracy. In general, to describe the time evolution of this new state we would require the state

$$|\psi(t)\rangle = a(t)|P^0\rangle + b(t)|\overline{P^0}\rangle + \sum_i c_i(t)|n_i\rangle\,, \tag{138}$$

where n_i are final states of the P^0 and $\overline{P^0}$ decays. However, we may study the mixing in this particle-antiparticle system separately from its subsequent decay if the following conditions are satisfied: $a(0),\, b(0) \neq 0$ and $c_i(0) = 0$; time scale larger than the typical strong-interaction scale; no interactions between final states (Weisskopf-Wigner approximation). In this way the neutral meson mixing is described by two-component wave function

$$\psi(t) = \begin{pmatrix} a(t) \\ b(t) \end{pmatrix} \tag{139}$$

evolving according to a Schrodinger equation

$$i\frac{d}{dt}\psi(t)\rangle = \underbrace{\left(M - \frac{i}{2}\Gamma\right)}_{H}\psi(t) = \begin{pmatrix} M_{11} - \frac{i}{2}\Gamma_{11} & M_{12} - \frac{i}{2}\Gamma_{12} \\ M_{21} - \frac{i}{2}\Gamma_{21} & M_{22} - \frac{i}{2}\Gamma_{22} \end{pmatrix}\psi(t)\,, \tag{140}$$

with t the proper time, H a 2×2 matrix written in the $P^0 - \overline{P^0}$ rest frame and M, Γ its Hermitian parts. The meson flavour basis $\left\{|P^0\rangle,\, |\overline{P^0}\rangle\right\}$ satisfies the following relations:

– **Orthogonality:** $\langle P^0|\overline{P^0}\rangle = \langle\overline{P^0}|P^0\rangle = 0$ and $\langle P^0|P^0\rangle = \langle\overline{P^0}|\overline{P^0}\rangle = 1$.

– **Completeness:** $|P^0\rangle\langle P^0| + |\overline{P^0}\rangle\langle\overline{P^0}| = 1$.

– **Effective Hamiltonian decomposition:** $\mathcal{H} = \left(|P^0\rangle, \quad |\overline{P^0}\rangle\right) \mathbf{H} \begin{pmatrix}\langle P^0| \\ \langle\overline{P^0}|\end{pmatrix}$

In terms of the total Hamiltonian

$$\mathcal{H} = \overbrace{\mathcal{H}_{\text{QCD}} + \mathcal{H}_{\text{QED}}}^{\textbf{CP}} + \overbrace{\mathcal{H}_{\text{EW}}}^{\textbf{CPV}} \tag{141}$$

we can have the usual perturbation expansion, up to second order,

$$\left(M - \frac{i}{2}\Gamma\right)_{ij} = \langle i|\mathcal{H}|j\rangle + \sum_n \frac{\langle i|\mathcal{H}|n\rangle\langle n|\mathcal{H}|j\rangle}{m_0 - (E_n - i\epsilon)} \tag{142}$$

where i and j can be K^0 or $\overline{K^0}$ and $|n\rangle$ any eigenstate of $\mathcal{H}_{\text{QCD}} + \mathcal{H}_{\text{QED}}$ with eigenvalue E_n, but with $n \neq K^0$, $\overline{K^0}$. Using the identity

$$\frac{1}{m_0 - (E_n - i\epsilon)} = P\frac{1}{m_0 - E_n} - i\pi\delta(m_0 - E_n) \tag{143}$$

we can find the Hermitian matrices M and Γ up to second order in perturbation theory. They are given by

$$\begin{aligned} M_{ij} &= \overbrace{\langle i|\mathcal{H}|j\rangle}^{m_0\delta_{ij} + \langle i|\mathcal{H}_{\text{EW}}|j\rangle} + \sum_n P\frac{\langle i|\mathcal{H}_{\text{EW}}|n\rangle\langle n|\mathcal{H}_{\text{EW}}|j\rangle}{m_0 - E_n}\,, \\ \Gamma_{ij} &= 2\pi\sum_n \delta(m_0 - E_n)\langle i|\mathcal{H}_{\text{EW}}|n\rangle\langle n|\mathcal{H}_{\text{EW}}|j\rangle\,, \end{aligned} \tag{144}$$

with P projecting out the principal part. The general CP transformation of the states is given by

$$\mathcal{CP}|P^0(\vec{p})\rangle = -e^{i\xi}|\overline{P^0}(-\vec{p})\rangle \quad \text{and} \quad \mathcal{CP}|\overline{P^0}(\vec{p})\rangle = -e^{-i\xi}|P^0(-\vec{p})\rangle\,. \tag{145}$$

We then see that the CP-invariant combinations are given by

$$|P_1\rangle = \frac{1}{\sqrt{2}}\left(|P^0\rangle - e^{i\xi}|\overline{P^0}\rangle\right), \quad |P_2\rangle = \frac{1}{\sqrt{2}}\left(|P^0\rangle + e^{i\xi}|\overline{P^0}\rangle\right), \tag{146}$$

in such a way that

$$\mathcal{CP}|P_1\rangle = |P_1\rangle \quad \text{and} \quad \mathcal{CP}|P_2\rangle = -|P_2\rangle\,. \tag{147}$$

Requesting CP invariance is equivalent to the Hamiltonian condition $\mathcal{H} = (\mathcal{CP})\mathcal{H}(\mathcal{CP})^\dagger$. This in turns imply $H_{12} = e^{-2i\xi}H_{21}$ and $H_{11} = H_{22}$. Note that ξ is a spurious phase without any physical relevance, therefore, we conclude that the phases of H_{12} and H_{21} also lack meaning. We can then summarize, in Table 6, the physical conditions given the present discrete symmetries. In these notes we are interested in $\mathcal{CPT}$-invariant theories.[2] As a result, the matrix responsible by the evolution of our system is given by

$$H = \begin{pmatrix} M_{11} - \frac{i}{2}\Gamma_{11} & M_{12} - \frac{i}{2}\Gamma_{12} \\ M_{12}^* - \frac{i}{2}\Gamma_{12} & M_{11} - \frac{i}{2}\Gamma_{11}\end{pmatrix}. \tag{148}$$

If CP was a symmetry of the system, i.e. $[\mathcal{CP}, \mathcal{H}] = 0$, the states $|P_{1,2}\rangle$ would be the true eigenstates of Eq. (140). The presence of CP-violating terms will destroy this result, in order to see this we go to the mass basis. The time evolution in Eq. (140) becomes trivial in the mass basis where the Hamiltonian H

[2] the general framework can be found in [2], for example.

Table 6: Constrains on the mixing matrix when the system respect some or no discrete symmetries.

Conservation	Constraints
$\mathcal{CPT}$	$H_{11} = H_{22}$ $(M_{11} = M_{22}$ and $\Gamma_{11} = \Gamma_{22})$
$\mathcal{CP}$	$H_{11} = H_{22}$ and $\lvert H_{12}\rvert = \lvert H_{21}\rvert$
$\mathcal{T}$	$\lvert H_{12}\rvert = \lvert H_{21}\rvert$
None	H is general

is diagonal. The complex eigenvalues ($\mu_{L,H}$) and corresponding eigenvectors ($|P_{L,H}\rangle$) of H are given by (using the phase convention $\xi = 0$)

Eigenvalues: **Eigenvectors:**

$$\left[\begin{array}{l} M_{11} - \frac{i}{2}\Gamma_{11} - pq = \mu_L = m_L - \frac{i}{2}\Gamma_L\,, \\ \\ M_{11} - \frac{i}{2}\Gamma_{11} + pq = \mu_H = m_H - \frac{i}{2}\Gamma_H \end{array}\right. \quad \left[\begin{array}{l} |P_L\rangle = \frac{1}{\sqrt{|p|^2+|q|^2}}\left(p|P^0\rangle - q|\overline{P^0}\rangle\right)\,, \\ \\ |P_H\rangle = \frac{1}{\sqrt{|p|^2+|q|^2}}\left(p|P^0\rangle + q|\overline{P^0}\rangle\right) \end{array}\right. \tag{149}$$

where

$$p^2 = M_{12} - \frac{i}{2}\Gamma_{12}\,, \quad q^2 = M_{12}^* - \frac{i}{2}\Gamma_{12}^*\,. \tag{150}$$

Note that $m_{H,L}$ and $\Gamma_{H,L}$ are not eigenvalues of M and Γ but, nevertheless, satisfy the relations $\mathrm{Tr}\, M = m_H + m_L = 2M_{11}$ and $\mathrm{Tr}\,\Gamma = \Gamma_H + \Gamma_L = 2\Gamma_{11}$. They can also be written as

$$\begin{aligned} m_L =& M_{11} - \mathrm{Re}\, pq\,, \quad m_H = M_{11} + \mathrm{Re}\, pq\,, \\ \Gamma_L =& \Gamma_{11} + 2\mathrm{Im}\, pq\,, \quad \Gamma_H = \Gamma_{11} - 2\mathrm{Im}\, pq\,. \end{aligned} \tag{151}$$

We are using a convention in which $\Delta m = m_H - m_L > 0$. It is also convenient to define

$$\mu \equiv \frac{\mu_H + \mu_L}{2} \equiv m - \frac{i}{2}\Gamma\,, \quad \Delta\mu \equiv \mu_H - \mu_L \equiv \Delta m - \frac{i}{2}\Delta\Gamma\,. \tag{152}$$

with

$$\begin{aligned} \Delta m =& m_H - m_L = 2\mathrm{Re}\, pq\,, \quad m = \frac{m_H + m_L}{2} = M_{11}\,, \\ \Delta\Gamma =& \Gamma_H - \Gamma_L = -4\mathrm{Im}\, pq\,, \quad \Gamma = \frac{\Gamma_H + \Gamma_L}{2} = \Gamma_{11}\,. \end{aligned} \tag{153}$$

The relation between these parameters and the elements of H in the flavour basis can be found through the diagonalization procedure, leading to

$$\mu = H_{11} = H_{22}\,, \quad \Delta\mu = 2\sqrt{H_{12}H_{21}}\,, \quad \frac{q}{p} = \sqrt{\frac{H_{21}}{H_{12}}} = \frac{2H_{21}}{\Delta\mu} \tag{154}$$

Which in a more familiar form can be written as

$$\begin{aligned} &(\Delta m)^2 - \frac{1}{4}(\Delta\Gamma)^2 = 4|M_{12}|^2 - |\Gamma_{12}|^2\,, \quad (\Delta m)(\Delta\Gamma) = 4\mathrm{Re}\left(M_{12}^*\Gamma_{12}\right)\,, \\ &\frac{1-\bar{\epsilon}}{1+\bar{\epsilon}} = \frac{q}{p} = \sqrt{\frac{M_{12}^* - \frac{i}{2}\Gamma_{12}^*}{M_{12} - \frac{i}{2}\Gamma_{12}}} = \frac{2M_{12}^* - i\Gamma_{12}^*}{\Delta m - \frac{i}{2}\Delta\Gamma} = \frac{\Delta m - \frac{i}{2}\Delta\Gamma}{2M_{12} - i\Gamma_{12}} \equiv re^{i\kappa}\,. \end{aligned} \tag{155}$$

The small complex parameter $\bar{\epsilon}$ depends on the phase convention chosen for the $P^0 - \overline{P^0}$ system. Therefore, as a spurious phase, it shall not be taken as a physical measure of CP violation. Nevertheless,

the quantities $\mathrm{Re}\,\bar{\epsilon}$ and r are independent of phase conventions. Therefore, departures of r from 1 are a measure of CP violation. If $r = 1$ ($\bar{\epsilon} = 0$) then $p = q$ and the mass eigenstates in Eq. (149) coincide with the CP eigenstates in Eq. (146). When this parameter is not 1 there is a small admixture of the CP eigenstates in the final (mass) eigenstates, i.e.

$$|P_L\rangle = \frac{1}{\sqrt{1+|\bar{\epsilon}|^2}}\left(|P_1\rangle + \bar{\epsilon}|P_2\rangle\right), \quad |P_H\rangle = \frac{1}{\sqrt{1+|\bar{\epsilon}|^2}}\left(|P_2\rangle + \bar{\epsilon}|P_1\rangle\right) \tag{156}$$

The physical observables measured in neutral meson oscillations can be parametrized by the dimensionless parameters

$$x = \frac{\Delta m}{\Gamma}, \quad y = \frac{\Delta\Gamma}{2\Gamma}, \quad r - 1 = \left|\frac{q}{p}\right| - 1\,. \tag{157}$$

On can check, after some algebra, that

$$\frac{|p|^2 - |q|^2}{|p|^2 + |q|^2} = \frac{1-r^2}{1+r^2} = \frac{\mathrm{Im}(M_{12}^*\Gamma_{12})}{|M_{12}|^2 + |\Gamma_{12}/2|^2 + \frac{1}{4}[(\Delta m)^2 + (\Delta\Gamma/2)^2]}\,, \tag{158}$$

which is actually the quantity which measures the non-orthogonality between $P_{L,H}$, i.e.

$$\langle P_H|P_L\rangle = \frac{1-r^2}{1+r^2} = \frac{2\mathrm{Re}\,\bar{\epsilon}}{1+|\bar{\epsilon}|^2}\,. \tag{159}$$

Concerning time evolution. For the $|P^0_{L,H}\rangle$ states the solutions is rather trivial

$$|P_{L,H}(t)\rangle = T_{L,H}(t)|P_{L,H}\rangle, \quad \text{with} \quad T_X(t) = e^{-i\mu_X t} = e^{-\Gamma_X t/2}e^{-im_X t}\,. \tag{160}$$

The states produced in strong interactions are the $|P^0\rangle$ and $|\overline{P^0}\rangle$. It turns then useful to look at the times evolutions for these states. Using Eq. (160) and Eq. (149), we find

$$\begin{aligned} |P^0(t)\rangle =& \frac{\sqrt{|p|^2+|q|^2}}{2p}\left[T_H(t)|P_H\rangle + T_L(t)|P_L\rangle\right], \\ |\overline{P^0}(t)\rangle =& \frac{\sqrt{|p|^2+|q|^2}}{2p}\left[T_H(t)|P_H\rangle - T_L(t)|P_L\rangle\right]. \end{aligned} \tag{161}$$

This form is useful for studies in the $K^0 - \overline{K^0}$ system. An alternative expression, useful in the $B^0 - \overline{B^0}$ system

$$|P^0(t)\rangle = f_+(t)|P^0\rangle + \frac{q}{p}f_-(t)|\overline{P^0}\rangle, \quad |\overline{P^0}(t)\rangle = f_+(t)|\overline{P^0}\rangle + \frac{p}{q}f_-(t)|P^0\rangle, \tag{162}$$

where

$$f_\pm(t) = \frac{T_H(t) \pm T_L(t)}{2} = \frac{1}{2}\left[e^{-im_H t}e^{-\Gamma_H t/2} \pm e^{-im_L t}e^{-\Gamma_L t/2}\right]. \tag{163}$$

One see right away that for $t = 0$ one has, for example, a pure $|P^0\rangle$ state, which as time evolves mixes with $|\overline{P^0}\rangle$. The probabilities of finding these states at later time are then given by

$$\begin{aligned} \mathcal{P}(P^0 \to P^0; t) =& \mathcal{P}(\overline{P^0} \to \overline{P^0}; t) = |f_+(t)|^2 = \frac{1}{2}\exp\left[-\frac{\Gamma t}{2}\right](\cos(\Delta m t) + \cosh(\Delta\Gamma/2)) \\ \mathcal{P}(P^0 \to \overline{P^0}; t) =& \left|\frac{q}{p}\right|^2 |f_-(t)|^2 = \frac{1}{2}\left|\frac{q}{p}\right|^2 \exp\left[-\frac{\Gamma t}{2}\right](-\cos(\Delta m t) + \cosh(\Delta\Gamma/2)) \\ \mathcal{P}(\overline{P^0} \to P^0; t) =& \left|\frac{p}{q}\right|^2 |f_-(t)|^2 = \frac{1}{2}\left|\frac{p}{q}\right|^2 \exp\left[-\frac{\Gamma t}{2}\right](-\cos(\Delta m t) + \cosh(\Delta\Gamma/2)) \end{aligned} \tag{164}$$

Note that several important aspects in meson oscillations were not covered here. For example, the existence of a reciprocal basis and its importance, this topic and many others can be found in [2, 9].

4.5 Neutral Meson Mixing: The $K^0 - \overline{K^0}$ and $B^0_{d,s} - \overline{B^0}_{d,s}$ systems

The general formalism for meson oscillations, shortly described in the previous section, can now be applied to the particular systems which we are interested in.

4.5.1 The $K^0 - \overline{K^0}$ system: $|K^0\rangle = |d\bar{s}\rangle$, $|\overline{K^0}\rangle = |\bar{d}s\rangle$

In this system instead of using the notation heavy (H) of light (L) for the mass eigenstates we change it to the standard notation of long (L) and short (S) life time particle. This means

$$|K_S\rangle \equiv |P_L\rangle \quad \text{and} \quad |K_L\rangle \equiv |P_H\rangle \,. \tag{165}$$

From the calculation of the $K_L - K_S$ mass difference, Gaillard and Lee [133] were able to estimate the value of the charm quark mass before its discovery. Also, kaon oscillation offers, within the Standard Model, a viable description of $\mathcal{CP}$ violation in $K_L \to \pi\pi$ decay.

In the kaon system we have

$$\tau_L \equiv \frac{1}{\Gamma_L} = 51.16 \pm 0.21\,\text{ps}\,, \quad \tau_S \equiv \frac{1}{\Gamma_S} = (0.8954 \pm 0.0004) \times 10^{-1}\text{ps} \tag{166}$$

$$m_K = 497.614 \pm 0.024\,\text{MeV}\,, \quad \Delta m_K = (3.484 \pm 0.006) \times 10^{-12}\,\text{MeV} \tag{167}$$

end up having to a good approximation

$$\Delta m_K \simeq 2|M_{12}| \simeq -\frac{1}{2}\Delta\Gamma_K \simeq |\Gamma_{12}|\,, \tag{168}$$

which lead us to

$$\frac{1-r^2}{1+r^2} \simeq \frac{1}{4}\text{Im}\left(\frac{\Gamma_{12}}{M_{12}}\right) \tag{169}$$

In order to relate $\bar{\epsilon}$ to measurable quantities we need to look at decays in the kaon system. The best channels to look at are the decay to pion. The pions are pseudo-scalars, which tell us that under the discrete symmetries C, P and T they transform in the same way as the bilinear $\overline{\psi}\gamma_5\chi$ in Tab. 2. Therefore, under CP we have

$$\begin{aligned}
&\textbf{One pion state:} \quad \mathcal{CP}|\pi^0\rangle = -|\pi^0\rangle\,, \\
&\textbf{Two pion state:} \quad \mathcal{CP}|\pi^0\pi^0\rangle = +|\pi^0\pi^0\rangle\,, \quad \mathcal{CP}|\pi^+\pi^-\rangle = +|\pi^+\pi^-\rangle\,, \\
&\textbf{Three pion state:} \quad \mathcal{CP}|\pi^0\pi^0\pi^0\rangle = -|\pi^0\pi^0\pi^0\rangle\,, \quad \mathcal{CP}|\pi^+\pi^-\pi^0\rangle = (-1)^l|\pi^+\pi^-\pi^0\rangle\,.
\end{aligned} \tag{170}$$

For the state $|\pi^+\pi^-\pi^0\rangle$ the relative angular momentum (l) between π^0 and $\pi^+\pi^-$ is relevant. We can then conclude, from the above properties, that a two pion final state is CP-even and a three pion final state (with zero angular momentum) CP-odd. The kaon decays to two or three pions can then be characterized as

$$\mathcal{CP}\textbf{ conserving:} \begin{bmatrix} K_S \to 2\pi \quad (\text{via } K_1) \\ \\ K_L \to 3\pi \quad (\text{via } K_2) \end{bmatrix} \qquad \mathcal{CP}\textbf{ violating:} \begin{bmatrix} K_S \to 3\pi \quad (\text{via } K_2) \\ \\ K_L \to 2\pi \quad (\text{via } K_1) \end{bmatrix} \tag{171}$$

This type of CP violation is called indirect since it comes from the presence of a small admixture of CP eigenstates in the final mass eigenstates, and not from a explicit breaking in the decay. We define the decay amplitudes:

$$\textbf{Decays:} \begin{bmatrix} \langle(\pi\pi)_{I=0}|\mathcal{H}|K^0\rangle = A_0 e^{i\delta_0} \\ \\ \langle(\pi\pi)_{I=2}|\mathcal{H}|K^0\rangle = A_2 e^{i\delta_2} \end{bmatrix}, \quad \textbf{CPT decays :} \begin{bmatrix} \langle(\pi\pi)_{I=0}|\mathcal{H}|\overline{K^0}\rangle = -A_0^* e^{i\delta_0} \\ \\ \langle(\pi\pi)_{I=2}|\mathcal{H}|\overline{K^0}\rangle = -A_2^* e^{i\delta_2} \end{bmatrix}. \tag{172}$$

Here δ_0 and δ_2 are the phase shifts where isospin quantum number $I = 0$ and $I = 2$ in $\pi\pi$ scattering. These are strong phases, and thus they do not change sign under CPT conjugation. These phases were factored out explicitly so that the phases of $A_{0,2}$ are all of weak nature

$$A_0 = |A_0|\text{exp}[i\phi_0]\,, \quad A_2 = |A_2|\text{exp}[i\phi_2]\,. \tag{173}$$

From the combination of Eq. (172) and Eq. (149) we get

$$\begin{aligned} A_0^{S,L} \equiv&\langle(\pi\pi)_{I=0}|\mathcal{H}_{EW}|K_{S,L}\rangle = \frac{pA_0 \pm qA_0^*}{\sqrt{|p|^2+|q|^2)}}\text{exp}[i\delta_0] = \frac{(1+\bar{\epsilon})A_0 \mp (1-\bar{\epsilon})A_0^*}{\sqrt{2(1+|\bar{\epsilon}|^2)}}\text{exp}[i\delta_0]\,, \\ A_2^{S,L} \equiv&\langle(\pi\pi)_{I=2}|\mathcal{H}_{EW}|K_{S,L}\rangle = \frac{pA_2 \pm qA_2^*}{\sqrt{|p|^2+|q|^2)}}\text{exp}[i\delta_0] = \frac{(1+\bar{\epsilon})A_2 \mp (1-\bar{\epsilon})A_2^*}{\sqrt{2(1+|\bar{\epsilon}|^2)}}\text{exp}[i\delta_2] \end{aligned} \tag{174}$$

Using the isotopic spin decomposition for the two pion states

$$\begin{aligned} \langle\pi^0\pi^0| &= \langle(\pi\pi)_{I=0}|\frac{1}{\sqrt{3}} - \langle(\pi\pi)_{I=2}|\sqrt{\frac{2}{3}}\,, \\ \frac{1}{\sqrt{2}}(\langle\pi^+\pi^-| + \langle\pi^-\pi^+|) &= \langle(\pi\pi)_{I=0}|\sqrt{\frac{2}{3}} + \langle(\pi\pi)_{I=2}|\frac{1}{\sqrt{3}}\,, \end{aligned} \tag{175}$$

where the charged pion state is correctly normalized, the transition amplitudes are defined as follow:

$$\begin{aligned} A(K_{S,L}\to\pi^0\pi^0) \equiv&\langle\pi^0\pi^0|\mathcal{H}_{EW}|K_{S,L}\rangle = \frac{1}{\sqrt{3}}A_0^{S,L} - \sqrt{\frac{2}{3}}A_2^{S,L}\,, \\ A(K_{S,L}\to\pi^+\pi^-) \equiv&\langle\pi^+\pi^-|\mathcal{H}_{EW}|K_{S,L}\rangle = \sqrt{\frac{2}{3}}A_0^{S,L} + \frac{1}{\sqrt{3}}A_2^{S,L}\,, \end{aligned} \tag{176}$$

and

$$\begin{aligned} A(K^0\to\pi^+\pi^-) \equiv&\langle\pi^+\pi^-|\mathcal{H}_{EW}|K^0\rangle = \frac{1}{\sqrt{3}}\left[\sqrt{2}A_0 + e^{i(\delta_2-\delta_0)}A_2\right] \\ A(\overline{K^0}\to\pi^+\pi^-) \equiv&\langle\pi^+\pi^-|\mathcal{H}_{EW}|\overline{K^0}\rangle = -\frac{1}{\sqrt{3}}\left[\sqrt{2}A_0^* + e^{i(\delta_2-\delta_0)}A_2^*\right] \\ A(K^0\to\pi^0\pi^0) \equiv&\langle\pi^0\pi^0|\mathcal{H}_{EW}|K^0\rangle = \frac{1}{\sqrt{3}}\left[A_0 - \sqrt{2}e^{i(\delta_2-\delta_0)}A_2\right] \\ A(\overline{K^0}\to\pi^0\pi^0) \equiv&\langle\pi^0\pi^0|\mathcal{H}_{EW}|\overline{K^0}\rangle = -\frac{1}{\sqrt{3}}\left[A_0^* - \sqrt{2}e^{i(\delta_2-\delta_0)}A_2^*\right]. \end{aligned} \tag{177}$$

Experimentally the decay of K_L to two-pion final state is observed and one can define useful quantities that measure this CP violation, i.e.

$$\begin{aligned} \eta_{00} =&\frac{A(K_L\to\pi^0\pi^0)}{A(K_S\to\pi^0\pi^0)} = \frac{A_0^L - \sqrt{2}A_2^L}{A_0^S - \sqrt{2}A_2^S} = \epsilon - \frac{2\epsilon'}{1-\sqrt{2}\omega}\,, \\ \eta_{+-} =&\frac{A(K_L\to\pi^+\pi^-)}{A(K_S\to\pi^+\pi^-)} = \frac{\sqrt{2}A_0^L + A_2^L}{\sqrt{2}A_0^S + A_2^S} = \epsilon + \frac{\epsilon'}{1+\omega/\sqrt{2}}\,, \end{aligned} \tag{178}$$

where $\omega \equiv \text{Re}[A_2/A_0]e^{i(\delta_2-\delta_0)}$. The experimental values for these quantities are [30]

$$\eta_{00} = (2.221 \pm 0.011)\times 10^{-3}\,\text{exp}[i(43.52\pm 0.06)^\circ]\,, \tag{179}$$

$$\eta_{+-} = (2.232 \pm 0.011)\times 10^{-3}\,,\text{exp}[i(43.51\pm 0.05)^\circ]\,, \tag{180}$$

showing how close these two quantities are. However, the fact that $\eta_{00} \neq \eta_{+-}$ is the source of CP violation in the kaon decay to two-pion final states. The parameter ϵ is the measure of indirect CP violation, which can be parametrized by amplitude ratio

$$\epsilon \equiv \frac{A_0^L}{A_0^S} \simeq \bar{\epsilon} + i\xi = \frac{e^{i\pi/4}}{\sqrt{2}\Delta m_K} \left(\text{Im}\, M_{12} + 2\xi \text{Re}\, M_{12}\right) , \quad \xi = \frac{\text{Im}[A_0]}{\text{Re}[A_0]} . \tag{181}$$

Both $\bar{\epsilon}$ and ξ have phase dependent conventions; however, since η_{+-} and η_{00} are experimental quantities ϵ is convention independent (similar to ϵ'). For direct CP violation parameter ϵ', where we have a direct transition of a CP-odd (even) term to a CP-even (odd), it is convenient to parametrize it through the following relation

$$\epsilon' = \frac{1}{\sqrt{2}} \left(\frac{A_2^L}{A_0^S} - \frac{A_2^S}{A_0^S} \frac{A_0^L}{A_0^S} \right) . \tag{182}$$

For small $\bar{\epsilon}$, i.e.$|\bar{\epsilon}| \ll 1$, we can then write

$$\epsilon \simeq \bar{\epsilon} + i \frac{\text{Im}[A_0]}{\text{Re}[A_0]} , \quad \epsilon' \simeq \frac{-ie^{i\Phi'}}{\sqrt{2}} \frac{\text{Re}[A_2]}{\text{Re}[A_0]} \left[\frac{\text{Im}[A_2]}{\text{Re}[A_2]} - \frac{\text{Im}[A_0]}{\text{Re}[A_0]} \right] . \tag{183}$$

It is possible by a choice of phase convention to set $\text{Im}[A_0] = 0$, known as Wu and Yang phase convention. The expressions are then simplified to

$$\textbf{In Wu-Yang phase convention:} \quad \begin{cases} \epsilon & \simeq \frac{1}{3}(2\eta_{+-} + \eta_{00}) \simeq \bar{\epsilon} \\ \\ \epsilon' & \simeq \frac{1}{3}(\eta_{+-} - \eta_{00}) \simeq \dfrac{e^{i\Phi'}}{\sqrt{2}} \dfrac{\text{Im}[A_2]}{\text{Im}[A_0]} \end{cases} , \tag{184}$$

where $\Phi' = \pi/2 + \delta_2 - \delta_0 \simeq \pi/4$. The parameter ϵ', which is only non-zero if there is CP violation in the decay amplitudes is proportional to the difference of η_{+-} and η_{00}, which almost cancel. A more practical quantity to evaluate ϵ' is the ratio given by

$$\text{Re}(\epsilon'/\epsilon) \simeq \frac{1}{6(1+\omega/\sqrt{2})} \left(1 - \left| \frac{\eta_{00}}{\eta_{+-}} \right|^2 \right) . \tag{185}$$

The parameter ω is small, i.e. $|\omega| \sim 1/25$, and often ignored. This quantity can be accurately measured on the rations $\Gamma(K_L \to \pi^0\pi^0)/\Gamma(K_L \to \pi^+\pi^-)$ and $\Gamma(K_S \to \pi^0\pi^0)/\Gamma(K_S \to \pi^+\pi^-)$, in terms of which

$$\left| \frac{\eta_{00}}{\eta_{+-}} \right|^2 = \frac{\Gamma(K_L \to \pi^0\pi^0)/\Gamma(K_L \to \pi^+\pi^-)}{\Gamma(K_S \to \pi^0\pi^0)/\Gamma(K_S \to \pi^+\pi^-)} . \tag{186}$$

From the fit to $K \to \pi\pi$ data we get [30]

$$|\epsilon| = (2.228 \pm 0.011) \times 10^{-3} , \quad \text{Re}[\epsilon'/\epsilon] = (1/65 \pm 0.26) \times 10^{-3} . \tag{187}$$

Another important observable is the CP asymmetry of time integrated semi-leptonic decay rates

$$\delta_L \equiv \frac{\Gamma(K_L \to \ell^+\nu_\ell\pi^-) - \Gamma(K_L \to \ell^-\bar{\nu}_\ell\pi^+)}{\Gamma(K_L \to \ell^+\nu_\ell\pi^-) + \Gamma(K_L \to \ell^-\bar{\nu}_\ell\pi^+)} = \frac{1 - \left|\frac{q}{p}\right|^2}{1 + \left|\frac{q}{p}\right|^2} = \frac{2\text{Re}[\bar{\epsilon}]}{1+|\bar{\epsilon}|^2} \longrightarrow \overbrace{\frac{2\text{Re}[\epsilon]}{1+|\epsilon|^2}}^{\textbf{Wu-Yang}} \tag{188}$$

This observable measure the orthogonality between K_L and K_S, see Eq. (158).

We can now shortly evaluate ϵ within the SM. The off-diagonal element M_{12} in the kaon system is given by

$$2m_K M_{12}^* = \langle \overline{K^0} | \mathcal{H}_{eff}^{\Delta S=2} | K^0 \rangle , \tag{189}$$

where the factor $2m_K$ is due to the normalization of external states. $|\mathcal{H}_{\text{eff}}^{\Delta S=2}$ is the effective Hamiltonian for the $\Delta S = 2$ transitions, this in lower order is given by the box diagrams in Fig. 2a. We can integrate out the heavy internal particles and run down to low energies with the renormalization group. By doing this we obtain the contact term

$$Q(\Delta S = 2) = (\bar{s}d)_{V-A}(\bar{s}d)_{V-A}\,. \tag{190}$$

The effective Hamiltonian, including leading and next-to-leading QCD corrections in the improved RGEs, for scales $\mu < \mu_c = \mathcal{O}(m_c)$ is given by

$$\begin{aligned}\mathcal{H}_{eff}^{\Delta S=2} =& \frac{G_F^2}{16\pi^2} M_W^2 \left[(V_{cs}^* V_{cd})^2 \eta_1 S_0(x_c) + (V_{ts}^* V_{td})^2 \eta_2 S_0(x_t) + 2(V_{cs}^* V_{cd})(V_{ts}^* V_{td})\eta_3 S_0(x_c, x_t)\right] \\ & \times [\alpha_s^{(3)}(\mu)]^{-2/9} \left[1 + \frac{\alpha_s^{(3)(\mu)}}{4\pi} J_3\right] Q(\Delta S = 2) + \text{h.c.}\end{aligned} \tag{191}$$

with $\alpha_s^{(3)}$ the strong coupling constant in an effective three flavour theory and $J_3 = 1.895$ in NDR scheme [6]. The S_0 loop functions are given by ($x_i = m_i^2/M_W^2$)

$$\begin{aligned} S_0(x_t) =& 2.39 \left(\frac{m_t}{167\,\text{GeV}}\right)^{1.52}, \quad S_0(x_c) = x_c\,, \\ S_0(x_c, x_t) =& x_c \left[\ln\frac{x_t}{x_c} - \frac{3x_t}{4(1-x_t)} - \frac{3x_t^2 \ln x_t}{4(1-x_t)^2}\right]. \end{aligned} \tag{192}$$

The factors $\eta_{1,2,3}$ are correction factors describing short distance QCD effects and at NLO read [6]: $\eta_1 = 1.38 \pm 0.20$, $\eta_2 = 0.57 \pm 0.01$, $\eta_3 = 0.47 \pm 0.04$. We can now take the matrix element of our contact interaction, the non-perturbative part of the calculation, we get

$$\langle \overline{K^0}|Q(\Delta S = 2)|K^0\rangle \equiv \frac{8}{3} B_K(\mu) F_K^2 m_K^2\,, \quad \hat{B}_K = B_K(\mu)[\alpha_s^{(3)}(\mu)]^{-2/9}\left[1 + \frac{\alpha_s^{(3)}(\mu)}{4\pi} J_3\right], \tag{193}$$

where $\hat{B}_K$ is a renormalization group invariant parameter and $F_K = 160$ MeV is the kaon decay constant. We finally find the matrix element to be

$$\begin{aligned} M_{12} = \frac{G_F^2}{12\pi^2} F_K^2 \hat{B}_K m_K M_W^2 \left[(V_{cs}^* V_{cd})^2 \eta_1 S_0(x_c) + (V_{ts}^* V_{td})^2 \eta_2 S_0(x_t)\right. \\ \left. + 2(V_{cs}^* V_{cd})(V_{ts}^* V_{td})\eta_3 S_0(x_c, x_t)\right]\,. \end{aligned} \tag{194}$$

Inserting this last result into Eq. (181) we obtain, in the Wu-Yang phase convention,

$$\epsilon \simeq C_\epsilon \hat{B}_K \text{Im}[V_{ts}^* V_{td}] \left\{\text{Re}[V_{cs}^* V_{cd}]\,[\eta_1 S_0(x_c) - \eta_3 S_0(x_c, x_t)] - \text{Re}[V_{ts}^* V_{td}]\eta_2 S_0(x_t)\right\} e^{i\pi/4}\,, \tag{195}$$

with

$$C_\epsilon = \frac{G_F^2 F_K^2 m_K M_W^2}{6\sqrt{2}\pi^2 \Delta m_K} \simeq 3.837 \times 10^4\,. \tag{196}$$

Corrections of the order $\text{Re}[V_{ts}^* V_{td}]/\text{Re}[V_{cs}^* V_{cd}] = \mathcal{O}(\lambda^4)$ have been neglected and we have used the unitary relation $\text{Im}[(V_{cs}^* V_{cd})^*] = \text{Im}[V_{ts}^* V_{td}]$. Using the standard CKM parametrization, Eq. (29), and comparing Eq. 195 with the experimental value Eq. (187) we can extract the CKM CP phase δ, important for the unitary triangle analysis.

The $K_L - K_S$ mass difference is now trivial to extract from Eqs. (194) and (168). Using the fact that $|V_{ts}^* V_{td}| \ll |V_{cs}^* V_{cd}|$, the charm-quark contribution in the loop dominates and we get

$$\Delta m_K \simeq \frac{G_F^2}{12\pi^2} F_K^2 \hat{B}_K m_K M_W^2 |V_{cs}^* V_{cd}|^2 S_0(x_c)\,. \tag{197}$$

4.5.2 The $B^0_{d,s} - \overline{B^0}_{d,s}$ system: $|B^0_d\rangle = |\bar{b}d\rangle$, $|\overline{B^0}_d\rangle = |b\bar{d}\rangle$, $|B^0_s\rangle = |\bar{b}s\rangle$, $|\overline{B^0}_s\rangle = |b\bar{s}\rangle$

Contrarily to the $K^0 - \overline{K^0}$ system, in the $B^0_{d,s} - \overline{B^0}_{d,s}$ system the long distance effects are very small $|\Gamma_{12}| \ll |M_{12}|$ (see discussion in [9]). Therefore, to leading order in $|\Gamma_{12}/M_{12}|$, we get

$$\Delta m_{B_q} = 2|M^q_{12}|\,, \quad \Delta\Gamma_{B_q} = 2\text{Re}(M^{q*}_{12}\Gamma^q_{12})/|M^q_{12}|\,, \quad \frac{q}{p} \simeq \frac{M^{q*}_{12}}{|M^q_{12}|}\left[1 - \frac{1}{2}\text{Im}\left(\frac{\Gamma^q_{12}}{M^q_{12}}\right)\right]\,, \tag{198}$$

with $q = d, s$ and the notation of H, L states given in the general discussion is kept here. In the B-system we have $|V^*_{td}V_{tb}| \sim |V^*_{cd}V_{cb}|$, however due to the quarks spectrum, i.e. $m_{u,c} \ll m_t$, the top quark contribution is now the one dominating.

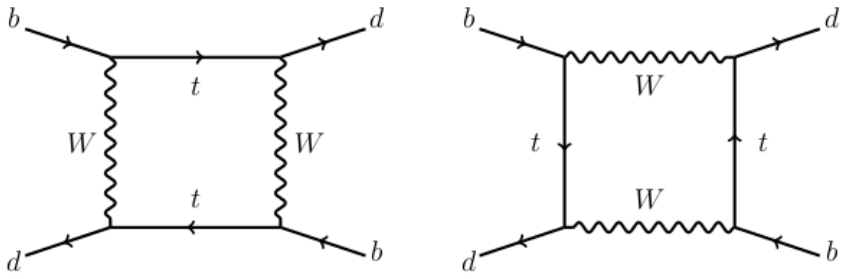


Fig. 18: Box diagram contributing to $B^0 = \overline{B^0}$ mixing

In a similar way as was done for the K-system, the off-diagonal element $M^{(q)}_{12}$ is given by

$$2m_{B_q}|M^{(q)}_{12}| = |\langle\overline{B^0}_q|\mathcal{H}^{\Delta B=2}_{\text{eff}}|B^0_q\rangle|\,. \tag{199}$$

The effective Hamiltonian , obtained from integrating out the top quark, is given by

$$\mathcal{H}^{\Delta B=2}_{\text{eff}} = \frac{G^2_F}{16\pi^2}M^2_W(V^*_{tb}V_{tq})^2\eta_B S_0(x_t)[\alpha^{(5)}_s(\mu_q)]^{-6/23}\left[1 + \frac{\alpha^{(5)}_s(\mu_q)}{4\pi}J_b\right]Q(\Delta B = 2) + \text{h.c.}\,, \tag{200}$$

with $\mu_q = \mathcal{O}(m_q)$ and $J_5 = 1.627$. The contact term is given by

$$Q(\Delta B = 2) = (\bar{b}q)_{V-A}(\bar{b}q)_{V-A}\,. \tag{201}$$

Taking the matrix element we get, in an analogous ways as for the K system,

$$\langle\overline{B^0}_q|Q(\Delta B = 2)|B^0_q\rangle \equiv \frac{8}{3}B_{B_q}(\mu)F^2_{B_q}m^2_{B_q}\,, \quad \hat{B}_{B_q} = B_{B_q}(\mu)[\alpha^{(5)}_s(\mu_q)]^{-6/23}\left[1 + \frac{\alpha^{(5)}_s(\mu_q)}{4\pi}J_b\right]\,, \tag{202}$$

with F_{B_q} the decay constant for B_q. Using Eq. (199) and the first relation in Eq. (198) one gets

$$\Delta m_{B_q} \simeq \frac{G^2_F}{6\pi^2}\eta_B m_{B_q}\hat{B}_{B_q}F^2_{B_q}M_W S_0(x_t)|V_{tq}|^2\,. \tag{203}$$

This relation for the mass difference in important in the standard analysis of the unitary triangle.

5 Flavour Physics Beyond the SM

CP violation in the SM comes from the flavour sector. However, CP violation observed so far is too small by a factor of 10^{-16} to explain the absence of anti-matter, which means that physics beyond the SM (BSM) must exist. Therefore, a right question wouldn't be whether BSM exist or not, but at which

scale it will show up. For particle physicists, there are also two different reasons hinting us that surprises might be awaiting to be discovered by at around TeV scale.

The first reason is coming from so-called 'the fine-tuning/hierarchy problem, which is related to the lightness of the Higgs particle compared to a arbitrarily high scale (below PLACK scale). The recently discovered Higgs particle, which is the only missing piece of the Standard Model (SM), may be the first fundamental scalar particles we have discovered. It is employed for the electroweak symmetry breaking (EWSB) and for generating masses for the fermions. While it explains why the weak force, unlike all other forces, is very short-ranged, it also provide us a problem. In order to obtain the observed $\sim$ 125GeV, which is far much smaller than the size of quantum corrections from seemingly unrelated forces, a miraculous fine-tuning has to be invoked. However, this 'naturalness' problem can be solved, if new physics exists beyond the Higgs particle. And the corresponding new physics and new particles are predicted to be observed in the scale of EWSB.

The other reason is coming from cosmology. According to the standard model of cosmology, which is now well established, some twenty percent of the energy of the universe comes from matter that does not shine (that is, electromagnetically neutral), but is much more massive than neutrinos. There are no candidates among particles in the SM for this type of matter, so called "dark matter (DM)". The cosmological and astrophysical observations suggest us that the mass of the DM particles is light enough to be produced and observed at the TeV scale.

In a general picture of physics beyond the SM one can see the amplitude of a given process being described in the form

$$A(\text{in} \to \text{out}) \simeq A_0 \left[\frac{C_{SM}}{M_W^2} + \frac{C_{NP}}{\Lambda_{NP}^2} \right] . \tag{204}$$

The coefficients $C_{SM(NP)}$ will the depend of the process and SM extension. However, we can see that flavour physics can place strong constraints on new physics even beyond the LHC reach. In scenarios where new physics does not respect the SM symmetries or breaking pattern, the coefficients tend to be hierarchical $C_{SM} \ll C_{NP}$, allowing to probe large scales.

For example, in the SM there are only two $|\Delta F| = 2$ operators entering in $K^0 - \bar{K}^0$ and $B^0 - \bar{B}^0$ mixing, see Sec. 4. A common feature in NP flavour models is the presence of additional four-quark operators, which change the flavour number by two units. Those interactions can place a strong bounds on the NP scale. Without specifying its origin we can typically describe them through the effective Lagrangian

$$\mathcal{L}_{NP}^{|\Delta F|=2} = \frac{1}{\Lambda^2} \sum_{i=1}^{5} c_i^{q_\alpha q_\beta} \mathcal{Q}_i^{q_\alpha q_\beta} + \frac{1}{\Lambda^2} \sum_{i=1}^{3} \tilde{c}_i^{q_\alpha q_\beta} \tilde{\mathcal{Q}}_i^{q_\alpha q_\beta} \tag{205}$$

with the dimension six $|\Delta F| = 2$ operators given by [44]

$$\begin{array}{lll}
\mathcal{Q}_1^{q_\alpha q_\beta} = \left(\overline{q}_{\beta L}\gamma_\mu q_{\alpha L}\right)\left(\overline{q}_{\beta L}\gamma_\mu q_{\alpha L}\right) , & \tilde{\mathcal{Q}}_1^{q_\alpha q_\beta} = \left(\overline{q}_{\beta R}\gamma_\mu q_{\alpha R}\right)\left(\overline{q}_{\beta R}\gamma_\mu q_{\alpha R}\right) , \\
\mathcal{Q}_2^{q_\alpha q_\beta} = \left(\overline{q}_{\beta R} q_{\alpha L}\right)\left(\overline{q}_{\beta R} q_{\alpha L}\right) , & \tilde{\mathcal{Q}}_2^{q_\alpha q_\beta} = \left(\overline{q}_{\beta L} q_{\alpha R}\right)\left(\overline{q}_{\beta L} q_{\alpha R}\right) , \\
\mathcal{Q}_3^{q_\alpha q_\beta} = \overline{q}^a_{\beta R} q^b_{\alpha L} \overline{q}^b_{\beta R} q^a_{\alpha L} , & \tilde{\mathcal{Q}}_3^{q_\alpha q_\beta} = \overline{q}^a_{\beta L} q^b_{\alpha R} \overline{q}^b_{\beta L} q^a_{\alpha R} , \\
\mathcal{Q}_4^{q_\alpha q_\beta} = \left(\overline{q}_{\beta R} q_{\alpha L}\right)\left(\overline{q}_{\beta L} q_{\alpha R}\right) , & \\
\mathcal{Q}_5^{q_\alpha q_\beta} = \overline{q}^a_{\beta R} q^b_{\alpha L} \overline{q}^b_{\beta L} q^a_{\alpha R} . &
\end{array} \tag{206}$$

Table 7 summarizes the bounds on the new physics scale or Wilson coefficient. As seen in Table 7 new physics scale tends to be pushed to very high scales (several orders above the TeV scale) due to flavour constraints. Saying it in other way, in order to have new physics at the TeV scale we need it to have specific flavour structure not so different from that of the SM at low energies. The quest for viable new physics models is known as "New Physics flavour problem". In this section we will look at some extensions and their confrontation with flavour observables.

Table 7: Summary of the most relevant bounds on $d = 6$ four-quark flavour operators. Taken from [42]

Operator	Bounds on Λ in TeV ($c_i^{\mathbf{NP}} = 1$)		Bounds on $c_i^{\mathbf{NP}}$ ($\Lambda = 1$ TeV)		Observable
	Re	Im	Re	Im	
$(\overline{s_L}\gamma^\mu d_L)^2$	9.8×10^2	1.6×10^4	9.0×10^{-7}	3.4×10^{-9}	Δm_K; ϵ_K
$(\overline{s_R}d_L)(\overline{s_L}d_R)$	1.8×10^4	3.2×10^5	6.9×10^{-9}	2.6×10^{-11}	
$(\overline{c_L}\gamma^\mu u_L)^2$	1.2×10^3	2.9×10^3	5.6×10^{-7}	1.0×10^{-7}	Δm_D; $\lvert q/p\rvert$, ϕ_D
$(\overline{c_R}u_L)(\overline{c_L}u_R)$	6.2×10^3	1.5×10^4	5.7×10^{-8}	1.1×10^{-8}	
$(\overline{b_L}\gamma^\mu d_L)^2$	6.6×10^2	9.3×10^2	2.3×10^{-6}	1.1×10^{-6}	Δm_{B_d}; $S_{\psi K_S}$
$(\overline{b_R}d_L)(\overline{b_L}d_R)$	2.5×10^3	3.6×10^3	3.9×10^{-7}	1.9×10^{-7}	
$(\overline{b_L}\gamma^\mu s_L)^2$	1.4×10^2	2.5×10^2	5.0×10^{-5}	1.7×10^{-5}	Δm_{B_s}; $S_{\psi\phi}$
$(\overline{b_R}s_L)(\overline{b_L}s_R)$	4.8×10^2	8.3×10^2	8.8×10^{-6}	2.9×10^{-6}	

5.1 Minimal flavour Violation hypothesis

One popular solution to the flavour puzzle is the minimal flavour violation (MFV) hypothesis [43]. The MFV is not a model, but a simple framework for the flavour structure on new physics seen from and effective field theory point of view. The main assumptions are:

- No new operators beyond those present in the SM;
- All flavour changing transitions are governed by CKM, i.e. no new complex phases beyond those present in the SM

$$A(\text{in} \to \text{out}) \propto \lambda^i_{CKM} \underbrace{(F^i_{SM} + F^i_{NP})}_{\text{real}} . \tag{207}$$

In the SM the CKM is the only source of flavour violation and is approximately a unit matrix. The SM has no flavour changing neutral currents at tree level, and in this way CKM-induced flavour change interactions are guarantee to be small. If new physics is flavour-diagonal such that all the flavour-violation goes through the CKM, then we are guaranteed to have small effects. Therefore, just like in the SM, Yukawa couplings are the only sources of flavour symmetry breaking in physics beyond the SM. In MFV we then have a CKM and GIM suppression working in a similar way to the SM, allowing and EFT-like approach.

The effective approach of MFV takes into account the larger flavour group in the SM when the Yukawa intersections are absent, see Eq. (12). This symmetry is explicitly broken in the presence of the Yukawa terms, but we can formally restore it by promoting the Yukawa matrices to be spurions (appropriate dimensionless auxiliary fields), which transform under the flavour group in the appropriate way to make it invariant (see Fig. 19).

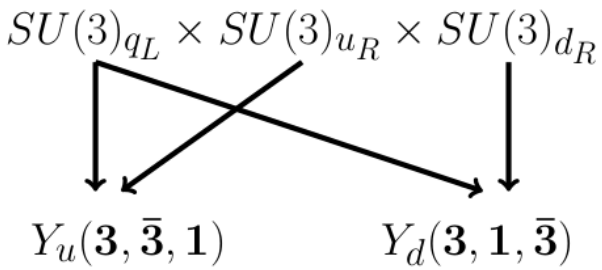


Fig. 19: Global flavour symmetry and spurious fields transformations

Using the $SU(3)_q^3 \times SU(3)_l^2$ symmetry, we can rotate the background values of the auxiliary field Y, as we did in Eq. (11),

$$Y_d = \lambda_d\,, \quad Y_u = V_{CKM}^\dagger \lambda_u\,, \quad Y_\ell = \lambda_\ell\,. \tag{208}$$

MFV requires that the dynamics of flavour violation is completely determined by the structure of the ordinary Yukawa couplings. In particular, all $\mathcal{CP}$ violation effects originates from the CKM phase. From the hierarchical structure of the Yukawa matrix, i.e. only top Yukawa is large, we can define the new physics flavour coupling

$$(\lambda_{FC})_{ij} = \begin{cases} (Y_u Y_u^\dagger)_{ij} \simeq y_t^2 V_{3i}^* V_{3j}\,, & i \neq j \\ 0\,, & i = j \end{cases} \tag{209}$$

The basic building blocks of FCNC operators are

Table 8: Relevant $d = 6$ MFV flavour operators and their bounds on new physics. Taken from [45].

MFV $d = 6$ operator	**Observables**	**Λ [TeV]**
$\frac{1}{2}(\overline{q_L}\lambda_{FC}\gamma_\mu q_L)^2$	ϵ_K, Δm_{D_d}	5.9
$\phi^\dagger(\overline{d_R}\lambda_d\lambda_{FC}\sigma_{\mu\nu}q_L)(eF^{\mu\nu})$	$B \to X_s\gamma$, $B \to X_s\ell^+\ell^-$	6.1
$\phi^\dagger(\overline{d_R}\lambda_d\lambda_{FC}\sigma_{\mu\nu}T^a q_L)(eg_s G^{a\mu\nu})$	$B \to X_s\gamma$, $B \to X_s\ell^+\ell^-$	3.4
$(\overline{q_L}\lambda_F C\gamma_\mu q_L)(eD_\nu F^{\mu\nu})$	$B \to X_s\ell^+\ell^-$	1.5
$i(\overline{q_L}\lambda_{FC}\gamma_\mu q_L)\phi^\dagger D^\mu\phi$	$B \to X_s\ell^+\ell^-$, $B_s \to \mu^+\mu^-$	1.1
$i(\overline{q_L}\lambda_{FC}\gamma_\mu\tau^a q_L)\phi^\dagger\tau^a D^\mu\phi$	$B \to X_s\ell^+\ell^-$, $B_s \to \mu^+\mu^-$	1.1
$(\overline{q_L}\lambda_{FC}\gamma_\mu q_L)(\overline{\ell_L}\gamma^\mu\ell_L)$	$B \to X_s\ell^+\ell^-$, $B_s \to \mu^+\mu^-$	1.7
$(\overline{q_L}\lambda_{FC}\gamma_\mu\tau^a q_L)(\overline{\ell_L}\gamma^\mu\tau^a\ell_L)$	$B \to X_s\ell^+\ell^-$, $B_s \to \mu^+\mu^-$	1.7
$(\overline{q_L}\lambda_{FC}\gamma_\mu q_L)(\overline{e_R}\gamma^\mu e_R)$	$B \to X_s\ell^+\ell^-$, $B_s \to \mu^+\mu^-$	2.7

$$\overline{q_L}Y_u Y_u^\dagger q_L\,, \quad \overline{d_R}Y_D^\dagger Y_u Y_u^\dagger q_L\,, \quad \overline{d_R}Y_d^\dagger Y_u Y_u^\dagger Y_d d_R \tag{210}$$

expanding in powers of the off-diagonal CKM matrix elements and in powers of the small Yukawa couplings, such as

$$\overline{q_L}\lambda_{FC}q_L \quad \text{and} \quad \overline{d_R}\lambda_d\lambda_{FC}q_L \tag{211}$$

The MFV framework is general and can be implemented in a given BSM scenario, e.g. SUSY and composite Higgs models, resulting in reducing the cutoff scale (flavour bound) from $\mathcal{O}(1000)$ TeV to $\mathcal{O}(1)$ TeV, which in turn makes it a very predictive theory framework. Compared to SM, only the flavour-independent magnitude of the transition amplitudes can be modified. A fingerprint of this framework is the prediction $(\sin 2\beta)_{B\to\psi K_s} = (\sin 2\beta)_{K\to\pi\nu\bar{\nu}}$, which can be identified by experiments.

5.2 Partial compositeness

Partial compositeness is a completely different way of flavour protection mechanism [46]. The idea is to generate quark and lepton masses through linear couplings of the Standard Model fields to composite operators, i.e.

$$\Delta_L\overline{q_L}\mathcal{O}_R + \Delta_R^u\overline{u_R}\mathcal{O}_R^u + \Delta_R^d\overline{d_R}\mathcal{O}_R^d + \cdots\,, \tag{212}$$

where $\Delta_{L,R}$ are known as pre-Yukawa couplings and $\mathcal{O}_{L,R}$ are fermionic operators arising from the strong sector. The nice aspect of this linear coupling is that no relevant operator can be built out of $\mathcal{O}_{L,R}$, since both have a classical mass dimension of $5/2$. Also, the quadratic operators $\mathcal{O}_L\mathcal{O}_L$, $\mathcal{O}_R\mathcal{O}_R$ vanish due to spinor identities and $\mathcal{O}_L\mathcal{O}_R$ is forbidden by gauge invariance. Therefore, the lowest-dimension operators on can build out of the composite operators are $\mathcal{O}_L\partial\!\!\!/\mathcal{O}_L$ and $\mathcal{O}_R\partial\!\!\!/\mathcal{O}_R$, which have classical dimension six and therefore irrelevant.

The physical light fermions will then be a mixture of both elementary and composite states, known as partial compositeness,

$$|\psi_{phys}\rangle = \cos\theta|\psi_{elem}\rangle + \sin\theta|\psi_{comp}\rangle\,. \tag{213}$$

The flavour problem in theories with strong dynamics can be improved if partial compositeness is implemented.

$$m_\Psi \simeq g_\Psi f \quad \longrightarrow \quad y_{SM} \simeq \frac{\Delta_L \Delta_R}{m_\Psi} \tag{214}$$

Partial compositeness provide partial solutions to both flavour and hierarchy puzzles. Still, this is a partial solution since from the kaon system ϵ_K and ϵ'_K/ϵ_K one still needs some sort of alignment, at least in the down sector. On the other hand, in this framework we can have a naturally sizable non-standard contribution to Δa_{CP}. This approach can be an alternative to MFV.

5.3 B physics at the LHC

Rare decays based on the flavour transition $b \to s$ have for some time call the attention of the flavour community, as they can be sensitive probes of new physics [47, 48]:

hadronic: $B \to \phi K$, $B \to \eta' K$, $B_s \to \phi\phi$, $B \to K\pi$, $B_s \to KK$, $\cdots$
radiative: $B \to X_s\gamma$, $B \to K^*\gamma$, $B_s \to \phi\gamma$, $\cdots$
semi-leptonic: $B \to X_s\ell\ell$, $B \to K\ell\ell$, $B \to K^*\ell\ell$, $B_s \to \phi\ell\ell$, $\cdots$
leptonic: $B_s \to \mu\mu$
neutrino: $B \to K\nu\bar{\nu}$, $B \to K^*\nu\bar{\nu}$

The most relevant ones in order to constrain new physics in the LHC era are the leptonic, semi-leptonic and radiative exclusive decays.

Recently, the LHCb collaboration observed an excess in $B \to K^*\mu^+\mu^-$ decay [49] by measuring the angular observables with a minimal sensitivity to the choice of form factors [50]. This tension can be soften by the presence of new physics. One useful way to search for new physics that could induce these deviations is to look at the effective Hamiltonian relevant for this transition. From the complete list presented in Sec. 3, the current-current, QCD penguin and electroweak penguin operators ar typically dominated by the SM contribution at low energies and will only contribute to the considered observables though mixing with the dominant operators. This effect is therefore small. The chromomagnetic dipole operators, for leptonic and semi-leptonic decays enter only through mixing. Tensor operator do not appear in $d = 6$ operator expansion the the SM. Having this information we can write the relevant effective Hamiltonian

$$\mathcal{H}_{eff} = -\frac{G_F}{\sqrt{2}}\frac{\alpha}{\pi}V_{tb}V_{ts}^*\sum_i\left(C_i^\ell\mathcal{O}_i^\ell + C_i^{\prime\ell}\mathcal{O}_i^{\prime\ell}\right) \tag{215}$$

with α the fine structure constant and the operators considered are

$$\begin{array}{ll}
\mathcal{O}_7 = \frac{m_b}{e}(\bar{s}\sigma_{\mu\nu}P_R b)F^{\mu\nu}\,, & \mathcal{O}'_7 = \frac{m_b}{e}(\bar{s}\sigma_{\mu\nu}P_L b)F^{\mu\nu}\,, \\
\mathcal{O}_9^\ell = (\bar{s}\gamma_\mu P_L b)(\bar{\ell}\gamma^\mu\ell)\,, & \mathcal{O}_9^{\prime\ell} = (\bar{s}\gamma_\mu P_R b)(\bar{\ell}\gamma^\mu\ell)\,, \\
\mathcal{O}_{10}^\ell = (\bar{s}\gamma_\mu P_L b)(\bar{\ell}\gamma^\mu\gamma_5\ell)\,, & \mathcal{O}_{10}^{\prime\ell} = (\bar{s}\gamma_\mu P_R b)(\bar{\ell}\gamma^\mu\gamma_5\ell)\,, \\
\mathcal{O}_S^\ell = (\bar{s}P_R b)(\bar{\ell}\ell)\,, & \mathcal{O}_S^{\prime\ell} = (\bar{s}P_L b)(\bar{\ell}\ell)\,, \\
\mathcal{O}_P^\ell = (\bar{s}P_R b)(\bar{\ell}\gamma_5\ell)\,, & \mathcal{O}_P^{\prime\ell} = (\bar{s}P_L b)(\bar{\ell}\gamma_5\ell)\,.
\end{array} \tag{216}$$

The operators $\mathcal{O}_{7-10}$ have been listed before, they are just written in the L, R notation instead of V, A one. The scalar and pseudo-scalar operators were also added, even though their impact is small in the observables. The prime operators are not present in the SM expansion, they therefore correspond always to new physics effects.

The presence of new physics in the relevant observables can be tracked to the corresponding operators:

- $B \to K\mu^+\mu^-$: $C_7^{(\prime)}$, $C_9^{(\prime)}$, $C_{10}^{(\prime)}$
- $B \to K^*\mu^+\mu^-$: $C_7^{(\prime)}$, $C_9^{(\prime)}$, $C_{10}^{(\prime)}$
- $B \to K^*\gamma$: $C_7^{(\prime)}$
- $B \to \phi\mu^+\mu^-$: $C_{10}^{(\prime)}$, $C_{S,P}^{(\prime)}$
- Lepton-nonuniversality: $C_9^{(\prime)}$, $C_{10}^{(\prime)}$
- $B \to \mu^+\mu^-$: $C_{10}^{(\prime)}$, $C_{S,P}$

In $B \to K\mu^+\mu^-$, $B \to K^*\mu^+\mu^-$, and $B \to \phi\mu^+\mu^-$ the form factors and contributions of the hadronic weak Hamiltonian are the main theoretical challenges. Direct CP asymmetries in B decays can give a hint of new physics, specially in $B \to K^*\gamma$ since the B factories measurements and LHCb are so precise. However, new physics in this observable is proportional to the strong phase that appears as a sub-leading effect and is also plagued with many uncertainties.

Several global fits have been done [47, 48], under the assumption of new physics entering only through one operator or two real Wilson coefficients. These analysis tend to favour values $C_9^{NP} < 0$ in order to accommodate the recent anomalies. New physics entering through C_9 can also contribute to the meson mixing. B_s-mixing is in general the most constraining observable.

Lepton-nouniversility is also a power probe of new physics. In the SM the process $b \to s\ell\ell$ is lepton flavour universal. However, beyond the SM new flavour violating interactions can give substantial deviation form lepton-universality. Ratios of branching fractions, as well as double ratios can serve as a clean probe of new physics [52,53]. A big advantage of considering ratios is the automatic cancelling of several uncertainties. Recently, the LHCb collaborations has reported [51]

$$R_K^{LHCb} = 0.745^{+0.090}_{-0.074} \pm 0.036 \tag{217}$$

which shows a 2.6σ deviation form the SM prediction $R_K^{SM} \simeq 1 + \mathcal{O}(m_\mu^2/m_b^2)$ [52], in the dilepton invariant mass squared bin $1\,\text{GeV}^2 \leq q^2 < 6\,\text{GeV}^2$. The branching fractions rations of rare semi-leptonic B decays of dimuons over dielectrons arge given by [52]

$$R_H = \frac{\mathcal{B}(\overline{B} \to \overline{H}\mu\mu)}{\mathcal{B}(\overline{B} \to \overline{H}ee)} \simeq \begin{cases} 1 + \Delta_+ + \Sigma_+\,, & H = K \\ 1 + \Delta_- + \Sigma_-\,, & H = K_0(1430) \\ 1 + p(\Delta_- - \Delta_+ + \Sigma_- - \Sigma_+) + \Delta_+ + \Sigma_+\,, & H = K^* \\ 1 + \frac{1}{2}(\Delta_- + \Delta_+ + \Sigma_- + \Sigma_+)\,, & H = X_s \end{cases} \tag{218}$$

while the double ratios are defined as

$$X_H \equiv \frac{R_H}{R_K} \simeq \begin{cases} 1 + (\Delta_- - \Delta_+ + \Sigma_- - \Sigma_+)\,, & H = K_0(1430) \\ 1 + p(\Delta_- - \Delta_+ + \Sigma_- - \Sigma_+)\,, & H = K^* \\ 1 + \frac{1}{2}(\Delta_- - \Delta_+ + \Sigma_- - \Sigma_+)\,, & H = X_s \end{cases} \tag{219}$$

with

$$\Delta_\pm = 2\frac{\text{Re}\left(C_9^{SM}(C_9^{NP\mu} \pm C_9^{\prime\mu})^*\right) + \text{Re}\left(C_{10}^{SM}(C_{10}^{NP\mu} \pm C_{10}^{\prime\mu})^*\right)}{|C_9^{SM}|^2 + |C_{10}^{SM}|^2} - (\mu \to e) \tag{220}$$

the new physics contribution from the interference with the SM, and

$$\Sigma_\pm = \frac{|C_9^{NP\mu} \pm C_9^{\prime\mu}|^2 + |C_{10}^{NP\mu} \pm C_{10}^{\prime\mu}|^2}{|C_9^{SM}|^2 + |C_{10}^{SM}|^2} - (\mu \to e) \tag{221}$$

the pure new physics contribution. At the m_b scale we have for the SM Wilson coefficients $C_9^{SM} = -C_{10}^{SM} \simeq 4.2$. The factor p is the polarization faction and is close to 1 (it is exactly 1 at zero recoil). These expression are valid to a very good accuracy given the current experimental uncertainties. The double ratios are very useful tool for precision tests new physics. They are only sensitive to new physics coupled to right-handed quarks, and therefore can be seen as complementary to R_H.

Another clean probe of new physics in the B sector are the leptonic decays $B \to \ell\ell$. The model independent average time-integrated branching ratio for $\overline{B}_s \to \ell\ell$ decays is [54]

$$\frac{\mathcal{B}(\overline{B}_s \to \ell\ell)}{\mathcal{B}(\overline{B}_s \to \ell\ell)^{SM}} = \left|1 - 0.24(C_{10}^{NP\ell} - C_{10}^{\prime\ell}) - y_\ell C_{P-}^{\ell}\right|^2 + |y_\ell C_{S-}^{\ell}|^2 \,, \tag{222}$$

with $y_u = 7.7$, $y_e = (m_\mu/m_e)y_\mu = 1.6 \times 10^3$ and $C_{P,S-} = C_{P,S} - C'_{P,S}$. The current reported experimental value for $\overline{B}_s \to \mu^+\mu^-$ decays is [55]

$$\frac{\mathcal{B}(\overline{B}_s \to \mu^+\mu^-)^{exp}}{\mathcal{B}(\overline{B}_s \to \mu^+\mu^-)^{SM}} = 0.79 \pm 0.20 \,. \tag{223}$$

Being a purely leptonic final state, the theoretical prediction of these processes is very clean and serves as a good probe for NP.

6 Brief conclusions

We have presented a short overview in the topics of flavour and CP violation in and beyond the SM. The most relevant aspects can be summarized as:

- Often we have seen the indirect evidence of New particles in flavour physics before directly discovering them;
- The SM flavour sector has been tested with impressive and increasing precision;
- In the SM, fermions come in 3 generations of quarks and leptons; flavour physics is all about them;
- All flavour violation in the SM is from the CKM matrix;
- CPV in SM is small, and comes from flavour;
- We have developed non relativistic QM tools for meson mixing;
- We have schematically shown how to calculate hadronic observables;
- Theoretical tools to understand the underlying physics is important. For example, effective field theory allows separation of different scales (separation of calculable parts and nonperturbative parts);
- Any sensitivity to high scales (including to physics beyond the Standard Model) can be treated using perturbative methods;
- Flavour structure of New Physics has to be special in order to be compatible with TeV scale New Physics. A popular example is MFV, but other possibilities exist such a partial compositeness, etc;
- If new particles discovered, their flavour properties can teach us about the underlying structure of New Physics: masses (degeneracies), decay rates (flavour decomposition), cross sections;
- Flavour physics provide important clues to model building in the LHC era;
- LHC era is also a Flavour Precision era, and a lot of interesting measurements are coming, as we have already seen some tensions with SM.

Acknowledgements

We wish to thank Monika Blanke for useful suggestions while preparing the lectures. We would like to also thank all the organisers of the second AEPSHEP School for the very warm hospitality extended to us and for the very nice atmosphere in the School. This work was supported by the National Research Foundation of Korea grant MEST No. 2012R1A2A2A01045722.

References

[1] S. Pokorski, *Gauge Field Theories*, Cambridge, Uk: Univ. Pr. (1987) 394 P. (Cambridge Monographs On Mathematical Physics).

[2] G. C. Branco, L. Lavoura and J. P. Silva, *CP Violation*, Int. Ser. Monogr. Phys. **103** (1999) 1.

[3] I. I. Y. Bigi and A. I. Sanda, *CP violation*, Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol. **9** (2000) 1.

[4] E. Leader and E. Predazzi, *An Introduction to gauge theories and modern particle physics. Vol. 2: CP violation, QCD and hard processes*, Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol. **4** (1996) 1.

[5] G. Buchalla, A. J. Buras and M. E. Lautenbacher, *Weak decays beyond leading logarithms*, Rev. Mod. Phys. **68** (1996) 1125 [hep-ph/9512380].

[6] A. J. Buras, *Weak Hamiltonian, CP violation and rare decays*, hep-ph/9806471.

[7] A. J. Buras, *Flavor dynamics: CP violation and rare decays*, hep-ph/0101336.

[8] G. Buchalla, *Kaon and charm physics: Theory*, hep-ph/0103166.

[9] J. P. Silva, *Phenomenological aspects of CP violation*, hep-ph/0410351.

[10] Y. Nir, *CP violation in meson decays*, hep-ph/0510413.

[11] A. Hocker and Z. Ligeti, *CP violation and the CKM matrix*, Ann. Rev. Nucl. Part. Sci. **56** (2006) 501 [hep-ph/0605217].

[12] M. Antonelli, D. M. Asner, D. A. Bauer, T. G. Becher, M. Beneke, A. J. Bevan, M. Blanke and C. Bloise *et al.*, *Flavor Physics in the Quark Sector*, Phys. Rept. **494** (2010) 197 [arXiv:0907.5386 [hep-ph]].

[13] Y. Grossman, *Introduction to flavor physics*, arXiv:1006.3534 [hep-ph].

[14] Y. Nir, *Flavour physics and CP violation*, CERN Yellow Report CERN-2010-001, 279-314 [arXiv:1010.2666 [hep-ph]].

[15] G. Isidori, *Flavor physics and CP violation*, arXiv:1302.0661 [hep-ph].

[16] B. Grinstein, *TASI-2013 Lectures on Flavor Physics*, arXiv:1501.05283 [hep-ph].

[17] Z. Ligeti, *TASI Lectures on Flavor Physics*, arXiv:1502.01372 [hep-ph].

[18] O. Gedalia and G. Perez, arXiv:1005.3106 [hep-ph].

[19] H. Georgi, *An Effective Field Theory for Heavy Quarks at Low-energies*, Phys. Lett. B **240** (1990) 447.

[20] M. Neubert, *Heavy quark symmetry*, Phys. Rept. **245** (1994) 259 [hep-ph/9306320].

[21] T. Mannel, *Heavy quark effective field theory*, Rept. Prog. Phys. **60** (1997) 1113.

[22] A. V. Manohar and M. B. Wise, *Heavy quark physics*, Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol. **10** (2000) 1.

[23] M. Neubert, *Effective field theory and heavy quark physics*, hep-ph/0512222.

[24] J. C. Romao and J. P. Silva, *A resource for signs and Feynman diagrams of the Standard Model*, Int. J. Mod. Phys. A **27** (2012) 1230025 [arXiv:1209.6213 [hep-ph]].

[25] N. Cabibbo, *Unitary Symmetry and Leptonic Decays*, Phys. Rev. Lett. **10** (1963) 531.

[26] M. Kobayashi and T. Maskawa, *CP Violation in the Renormalizable Theory of Weak Interaction*, Prog. Theor. Phys. **49** (1973) 652.

[27] S. L. Glashow and S. Weinberg, *Natural Conservation Laws for Neutral Currents*, Phys. Rev. D **15** (1977) 1958.

[28] J. Bernabeu, G. C. Branco and M. Gronau, *Cp Restrictions On Quark Mass Matrices*, Phys. Lett. B **169** (1986) 243.

[29] C. Jarlskog, *Commutator of the Quark Mass Matrices in the Standard Electroweak Model and a Measure of Maximal CP Violation*, Phys. Rev. Lett. **55** (1985) 1039.

[30] K. A. Olive *et al.* [Particle Data Group Collaboration], *Review of Particle Physics*, Chin. Phys. C **38** (2014) 090001.

[31] L. Wolfenstein, *Parametrization of the Kobayashi-Maskawa Matrix*, Phys. Rev. Lett. **51** (1983) 1945.

[32] S. L. Glashow, J. Iliopoulos and L. Maiani, *Weak Interactions with Lepton-Hadron Symmetry*, Phys. Rev. D **2** (1970) 1285.

[33] G. Amoros, M. Beneke and M. Neubert, *Two loop anomalous dimension of the chromomagnetic moment of a heavy quark*, Phys. Lett. B **401** (1997) 81 [hep-ph/9701375].

[34] M. Neubert and V. Rieckert, *New approach to the universal form-factors in decays of heavy mesons*, Nucl. Phys. B **382** (1992) 97.

[35] A. D. Linde, *Inflationary Cosmology*, Lect. Notes Phys. **738** (2008) 1 [arXiv:0705.0164 [hep-th]].

[36] C. L. Bennett *et al.* [WMAP Collaboration], *Nine-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results*, Astrophys. J. Suppl. **208** (2013) 20 [arXiv:1212.5225 [astro-ph.CO]].

[37] A. D. Sakharov, *Violation of CP Invariance, c Asymmetry, and Baryon Asymmetry of the Universe*, Pisma Zh. Eksp. Teor. Fiz. **5** (1967) 32 [JETP Lett. **5** (1967) 24] [Sov. Phys. Usp. **34** (1991) 392] [Usp. Fiz. Nauk **161** (1991) 61].

[38] L. L. Chau, *Quark Mixing in Weak Interactions*, Phys. Rept. **95** (1983) 1.

[39] K. Kleinknecht, *Uncovering CP violation: Experimental clarification in the neutral K meson and B meson*, Springer Tracts Mod. Phys. **195** (2003) 1.

[40] Z. Ligeti, *Introduction to heavy meson decays and CP asymmetries*, eConf C **020805** (2002) L02 [hep-ph/0302031].

[41] M. K. Gaillard and B. W. Lee, *Rare Decay Modes of the K-Mesons in Gauge Theories*, Phys. Rev. D **10** (1974) 897.

[42] G. Isidori, Y. Nir and G. Perez, *Flavor Physics Constraints for Physics Beyond the Standard Model*, Ann. Rev. Nucl. Part. Sci. **60** (2010) 355 [arXiv:1002.0900 [hep-ph]].

[43] G. D'Ambrosio, G. F. Giudice, G. Isidori and A. Strumia, *Minimal flavor violation: An Effective field theory approach*, Nucl. Phys. B **645** (2002) 155 [hep-ph/0207036].

[44] Y. Grossman, Y. Nir, J. Thaler, T. Volansky and J. Zupan, *Probing minimal flavor violation at the LHC*, Phys. Rev. D **76** (2007) 096006 [arXiv:0706.1845 [hep-ph]].

[45] T. Hurth, G. Isidori, J. F. Kamenik and F. Mescia, *Constraints on New Physics in MFV models: A Model-independent analysis of $\Delta F = 1$ processes*, Nucl. Phys. B **808** (2009) 326 [arXiv:0807.5039 [hep-ph]].

[46] D. B. Kaplan, *Flavor at SSC energies: A New mechanism for dynamically generated fermion masses*, Nucl. Phys. B **365** (1991) 259.

[47] W. Altmannshofer and D. M. Straub, *State of new physics in $b \to s$ transitions*, arXiv:1411.3161 [hep-ph].

[48] W. Altmannshofer and D. M. Straub, *Implications of $b \to s$ measurements*, arXiv:1503.06199 [hep-ph].

[49] R. Aaij *et al.* [LHCb Collaboration], *Measurement of Form-Factor-Independent Observables in the Decay $B^0 \to K^{*0}\mu^+\mu^-$*, Phys. Rev. Lett. **111** (2013) 191801 [arXiv:1308.1707 [hep-ex]].

[50] S. Descotes-Genon, T. Hurth, J. Matias and J. Virto, *Optimizing the basis of $B \to K^*\ell^+\ell^-$ observables in the full kinematic range*, JHEP **1305** (2013) 137 [arXiv:1303.5794 [hep-ph]].

[51] R. Aaij *et al.* [LHCb Collaboration], *Test of lepton universality using $B^+ \to K^+\ell^+\ell^-$ decays*, Phys. Rev. Lett. **113** (2014) 151601 [arXiv:1406.6482 [hep-ex]].

[52] G. Hiller and M. Schmaltz, *Diagnosing lepton-nonuniversality in $b \to s\ell\ell$*, JHEP **1502** (2015) 055 [arXiv:1411.4773 [hep-ph]].

[53] G. Hiller and M. Schmaltz, *R_K and future $b \to s\ell\ell$ physics beyond the standard model opportunities*, Phys. Rev. D **90** (2014) 054014 [arXiv:1408.1627 [hep-ph]].

[54] A. J. Buras, R. Fleischer, J. Girrbach and R. Knegjens, *Probing New Physics with the $B_s \to \mu + \mu$-Time-Dependent Rate*, JHEP **1307** (2013) 77 [arXiv:1303.3820 [hep-ph]].

[55] CMS and LHCb Collaborations [CMS and LHCb Collaborations], *Combination of results on the rare decays $B^0_{(s)} \to \mu^+\mu^-$ from the CMS and LHCb experiments*, CMS-PAS-BPH-13-007, LHCb-CONF-2013-012, CERN-LHCb-CONF-2013-012.

Proceedings of the 2014 Asia–Europe–Pacific School of High-Energy Physics, Puri, India, 4–17 November 2014, edited by
M. Mulders and R. Godbole, CERN Yellow Reports: School Proceedings, Vol. 2/2017, CERN-2017-005-SP (CERN, Geneva, 2017)

Neutrino Physics

Z.Z. Xing
Institute of High Energy Physics and Theoretical Physics Center for Science Facilities,
Chinese Academy of Sciences, Beijing, China

Abstract
I give a theoretical overview of some basic properties of massive neutrinos in these lectures. Particular attention is paid to the origin of neutrino masses, the pattern of lepton flavor mixing, the feature of leptonic CP violation and the electromagnetic properties of massive neutrinos. I highlight the TeV seesaw mechanisms as a possible bridge between neutrino physics and collider physics in the era characterized by the Large Hadron Collider.

Keywords
Lectures; neutrino; particle physics; neutrino oscillations; mixing; standard model.

1 Finite Neutrino Masses

It is well known that the mass of an elementary particle represents its inertial energy when it exists at rest. Hence a massless particle has no way to exist at rest — instead, it must always move at the speed of light. A massive fermion (either lepton or quark) must exist in both left-handed and right-handed states, since the field operators responsible for the non-vanishing mass of a fermion have to be bilinear products of the spinor fields which flip the fermion's handedness or chirality.

The standard model (SM) of electroweak interactions contains three neutrinos $(\nu_e, \nu_\mu, \nu_\tau)$ which are purely left-handed and massless. In the SM the masslessness of the photon is guaranteed by the electromagnetic $U(1)_{\rm Q}$ gauge symmetry. Although the masslessness of three neutrinos corresponds to the lepton number conservation [1], the latter is an accidental symmetry rather than a fundamental symmetry of the SM. Hence many physicists strongly believed that neutrinos should be massive even long before some incontrovertible experimental evidence for massive neutrinos were accumulated. A good reason for this belief is that neutrinos are more natural to be massive than to be massless in some grand unified theories, such as the $\rm SO(10)$ theory, which try to unify electromagnetic, weak and strong interactions as well as leptons and quarks.

If neutrinos are massive and their masses are non-degenerate, it will in general be impossible to find a flavor basis in which the coincidence between flavor and mass eigenstates holds both for charged leptons (e, μ, τ) and for neutrinos $(\nu_e, \nu_\mu, \nu_\tau)$. In other words, the phenomenon of flavor mixing is naturally expected to appear between three charged leptons and three massive neutrinos, just like the phenomenon of flavor mixing between three up-type quarks (u, c, t) and three down-type quarks (d, s, b). If there exist irremovable complex phases in the Yukawa interactions, CP violation will naturally appear both in the quark sector and in the lepton sector.

[1] It is actually the $B-L$ symmetry that makes neutrinos exactly massless in the SM, where B = baryon number and L = lepton number. The reason is simply that a neutrino and an antineutrino have different values of $B-L$. Thus the naive argument for massless neutrinos is valid to all orders in perturbation and non-perturbation theories, if $B-L$ is an exact symmetry.

https://doi.org/10.23730/CYRSP-2017-002.177

1.1 Some preliminaries

To write out the mass term for three known neutrinos, let us make a minimal extension of the SM by introducing three right-handed neutrinos. Then we totally have six neutrino fields [2]:

$$\nu_{\rm L} = \begin{pmatrix} \nu_{e{\rm L}} \\ \nu_{\mu{\rm L}} \\ \nu_{\tau{\rm L}} \end{pmatrix} \; , \quad N_{\rm R} = \begin{pmatrix} N_{1{\rm R}} \\ N_{2{\rm R}} \\ N_{3{\rm R}} \end{pmatrix} \; , \tag{1}$$

where only the left-handed fields take part in the electroweak interactions. The charge-conjugate counterparts of $\nu_{\rm L}$ and $N_{\rm R}$ are defined as

$$(\nu_{\rm L})^c \equiv \mathcal{C}\overline{\nu_{\rm L}}^T \; , \quad (N_{\rm R})^c \equiv \mathcal{C}\overline{N_{\rm R}}^T \; ; \tag{2}$$

and accordingly,

$$\overline{(\nu_{\rm L})^c} = (\nu_{\rm L})^T \mathcal{C} \; , \quad \overline{(N_{\rm R})^c} = (N_{\rm R})^T \mathcal{C} \; , \tag{3}$$

where $\mathcal{C}$ denotes the charge-conjugation matrix and satisfies the conditions

$$\mathcal{C}\gamma_\mu^T \mathcal{C}^{-1} = -\gamma_\mu \; , \quad \mathcal{C}\gamma_5^T \mathcal{C}^{-1} = \gamma_5 \; , \quad \mathcal{C}^{-1} = \mathcal{C}^\dagger = \mathcal{C}^T = -\mathcal{C} \; . \tag{4}$$

It is easy to check that $P_{\rm L}(N_{\rm R})^c = (N_{\rm R})^c$ and $P_{\rm R}(\nu_{\rm L})^c = (\nu_{\rm L})^c$ hold; namely, $(\nu_{\rm L})^c = (\nu^c)_{\rm R}$ and $(N_{\rm R})^c = (N^c)_{\rm L}$ hold. Hence $(\nu_{\rm L})^c$ and $(N_{\rm R})^c$ are right- and left-handed fields, respectively. One may then use the neutrino fields $\nu_{\rm L}$, $N_{\rm R}$ and their charge-conjugate partners to write out the gauge-invariant and Lorentz-invariant neutrino mass terms.

In the SM the weak charged-current interactions of three active neutrinos are given by

$$\mathcal{L}_{\rm cc} = \frac{g}{\sqrt{2}} \overline{(e \; \mu \; \tau)_{\rm L}} \; \gamma^\mu \begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix}_{\rm L} W^-_\mu + {\rm h.c.} \; . \tag{5}$$

Without loss of generality, we choose the basis in which the mass eigenstates of three charged leptons are identified with their flavor eigenstates. If neutrinos have non-zero and non-degenerate masses, their flavor and mass eigenstates are in general not identical in the chosen basis. This mismatch signifies lepton flavor mixing.

1.2 Dirac neutrino masses

A Dirac neutrino is described by a four-component Dirac spinor $\nu = \nu_{\rm L} + N_{\rm R}$, whose left-handed and right-handed components are just $\nu_{\rm L}$ and $N_{\rm R}$. The Dirac neutrino mass term comes from the Yukawa interactions

$$-\mathcal{L}_{\rm Dirac} = \overline{\ell_{\rm L}} Y_\nu \tilde{H} N_{\rm R} + {\rm h.c.} \; , \tag{6}$$

where $\tilde{H} \equiv i\sigma_2 H^*$ with H being the SM Higgs doublet, and $\ell_{\rm L}$ denotes the left-handed lepton doublet. After spontaneous gauge symmetry breaking (i.e., $SU(2)_{\rm L} \times U(1)_{\rm Y} \to U(1)_{\rm Q}$), we obtain

$$-\mathcal{L}'_{\rm Dirac} = \overline{\nu_{\rm L}} M_{\rm D} N_{\rm R} + {\rm h.c.} \; , \tag{7}$$

where $M_{\rm D} = Y_\nu \langle H \rangle$ with $\langle H \rangle \simeq 174$ GeV being the vacuum expectation value of H. This mass matrix can be diagonalized by a bi-unitary transformation: $V^\dagger M_{\rm D} U = \widehat{M}_\nu \equiv {\rm Diag}\{m_1, m_2, m_3\}$ with m_i being the neutrino masses (for $i = 1, 2, 3$). After this diagonalization,

$$-\mathcal{L}'_{\rm Dirac} = \overline{\nu'_{\rm L}} \widehat{M}_\nu N'_{\rm R} + {\rm h.c.} \; , \tag{8}$$

[2] The left- and right-handed components of a fermion field $\psi(x)$ are denoted as $\psi_{\rm L}(x) = P_{\rm L}\psi(x)$ and $\psi_{\rm R}(x) = P_{\rm R}\psi(x)$, respectively, where $P_{\rm L} \equiv (1-\gamma_5)/2$ and $P_{\rm R} \equiv (1+\gamma_5)/2$ are the chiral projection operators. Note, however, that $\nu_{\rm L} = P_{\rm L}\nu_{\rm L}$ and $N_{\rm R} = P_{\rm R}N_{\rm R}$ are in general independent of each other.

where $\nu'_{\rm L} = V^\dagger \nu_{\rm L}$ and $N'_{\rm R} = U^\dagger N_{\rm R}$. Then the four-component Dirac spinor

$$\nu' = \nu'_{\rm L} + N'_{\rm R} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} \; , \tag{9}$$

which automatically satisfies $P_{\rm L}\nu' = \nu'_{\rm L}$ and $P_{\rm R}\nu' = N'_{\rm R}$, describes the mass eigenstates of three Dirac neutrinos. In other words,

$$-\mathcal{L}'_{\rm Dirac} = \overline{\nu'} \widehat{M}_\nu \nu' = \sum_{i=1}^{3} m_i \overline{\nu_i} \nu_i \; . \tag{10}$$

The kinetic term of Dirac neutrinos reads

$$\mathcal{L}_{\rm kinetic} = i\overline{\nu_{\rm L}} \gamma_\mu \partial^\mu \nu_{\rm L} + i\overline{N_{\rm R}} \gamma_\mu \partial^\mu N_{\rm R} = i\overline{\nu'} \gamma_\mu \partial^\mu \nu' = i \sum_{k=1}^{3} \overline{\nu_k} \gamma_\mu \partial^\mu \nu_k \; , \tag{11}$$

where $V^\dagger V = V V^\dagger = \mathbf{1}$ and $U^\dagger U = U U^\dagger = \mathbf{1}$ have been used.

Now we rewrite the weak charged-current interactions of three neutrinos in Eq. (5) in terms of their mass eigenstates $\nu'_{\rm L} = V^\dagger \nu_{\rm L}$ in the chosen basis where the flavor and mass eigenstates of three charged leptons are identical:

$$\mathcal{L}_{\rm cc} = \frac{g}{\sqrt{2}} \overline{(e \; \mu \; \tau)_{\rm L}} \; \gamma^\mu V \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix}_{\rm L} W^-_\mu + {\rm h.c.} \; . \tag{12}$$

The 3×3 unitary matrix V, which actually links the neutrino mass eigenstates (ν_1, ν_2, ν_3) to the neutrino flavor eigenstates $(\nu_e, \nu_\mu, \nu_\tau)$, just measures the phenomenon of neutrino mixing.

A salient feature of massive Dirac neutrinos is lepton number conservation. To see why massive Dirac neutrinos are lepton-number-conserving, we make the global phase transformations

$$l(x) \to e^{i\Phi} l(x) \; , \quad \nu'_{\rm L}(x) \to e^{i\Phi} \nu'_{\rm L}(x) \; , \quad N'_{\rm R}(x) \to e^{i\Phi} N'_{\rm R}(x) \; , \tag{13}$$

where l denotes the column vector of e, μ and τ fields, and Φ is an arbitrary spacetime-independent phase. As the mass term $\mathcal{L}'_{\rm Dirac}$, the kinetic term $\mathcal{L}_{\rm kinetic}$ and the charged-current interaction term $\mathcal{L}_{\rm cc}$ are all invariant under these transformations, the lepton number must be conserved for massive Dirac neutrinos. It is evident that lepton flavors are violated, unless $M_{\rm D}$ is diagonal or equivalently V is the identity matrix. In other words, lepton flavor mixing leads to lepton flavor violation, or vice versa.

For example, the decay mode $\pi^- \to \mu^- + \overline{\nu}_\mu$ preserves both the lepton number and lepton flavors. In contrast, $\mu^+ \to e^+ + \gamma$ preserves the lepton number but violates the lepton flavors. The observed phenomena of neutrino oscillations verify the existence of neutrino flavor violation. Note that the $0\nu2\beta$ decay $(A, Z) \to (A, Z+2) + 2e^-$ violates the lepton number. This process cannot take place if neutrinos are massive Dirac particles, but it may naturally happen if neutrinos are massive Majorana particles.

1.3 Majorana neutrino masses

The left-handed neutrino field $\nu_{\rm L}$ and its charge-conjugate counterpart $(\nu_{\rm L})^c$ can in principle form a neutrino mass term, as $(\nu_{\rm L})^c$ is actually right-handed. But this Majorana mass term is forbidden by the $SU(2)_{\rm L} \times U(1)_{\rm Y}$ gauge symmetry in the SM, which contains only one $SU(2)_{\rm L}$ Higgs doublet and preserves lepton number conservation. We shall show later that the introduction of an $SU(2)_{\rm L}$ Higgs triplet into the SM can accommodate such a neutrino mass term with gauge invariance. Here we ignore the details of the Higgs triplet models and focus on the Majorana neutrino mass term itself:

$$-\mathcal{L}'_{\rm Majorana} = \frac{1}{2} \overline{\nu_{\rm L}} M_{\rm L} (\nu_{\rm L})^c + {\rm h.c.} \; . \tag{14}$$

Note that the mass matrix $M_{\rm L}$ must be symmetric. Because the mass term is a Lorentz scalar whose transpose keeps unchanged, we have

$$\overline{\nu_L} M_{\rm L} (\nu_L)^c = \left[\overline{\nu_L} M_{\rm L} (\nu_L)^c\right]^T = -\overline{\nu_L} \mathcal{C}^T M_{\rm L}^T \overline{\nu_L}^T = \overline{\nu_L} M_{\rm L}^T (\nu_L)^c \; , \tag{15}$$

where a minus sign appears when interchanging two fermion field operators, and $\mathcal{C}^T = -\mathcal{C}$ has been used. Hence $M_{\rm L}^T = M_{\rm L}$ holds. This symmetric mass matrix can be diagonalized by the transformation $V^\dagger M_{\rm L} V^* = \widehat{M}_\nu \equiv {\rm Diag}\{m_1, m_2, m_3\}$, where V is a unitary matrix. After this, Eq. (14) becomes

$$-\mathcal{L}'_{\rm Majorana} = \frac{1}{2} \overline{\nu'_{\rm L}} \widehat{M}_\nu (\nu'_{\rm L})^c + {\rm h.c.} \; , \tag{16}$$

where $\nu'_{\rm L} = V^\dagger \nu_{\rm L}$ and $(\nu'_{\rm L})^c = \mathcal{C} \overline{\nu'_{\rm L}}^T$. Then the Majorana field

$$\nu' = \nu'_{\rm L} + (\nu'_{\rm L})^c = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} \; , \tag{17}$$

which certainly satisfies the Majorana condition $(\nu')^c = \nu'$, describes the mass eigenstates of three Majorana neutrinos. In other words,

$$-\mathcal{L}'_{\rm Majorana} = \frac{1}{2} \overline{\nu'} \widehat{M}_\nu \nu' = \frac{1}{2} \sum_{i=1}^{3} m_i \overline{\nu}_i \nu_i \; . \tag{18}$$

The kinetic term of Majorana neutrinos reads

$$\mathcal{L}_{\rm kinetic} = i \overline{\nu_{\rm L}} \gamma_\mu \partial^\mu \nu_{\rm L} = i \overline{\nu'_{\rm L}} \gamma_\mu \partial^\mu \nu'_{\rm L} = \frac{i}{2} \overline{\nu'} \gamma_\mu \partial^\mu \nu' = \frac{i}{2} \sum_{k=1}^{3} \overline{\nu_k} \gamma_\mu \partial^\mu \nu_k \; , \tag{19}$$

where we have used a generic relationship $\overline{(\psi_{\rm L})^c} \gamma_\mu \partial^\mu (\psi_{\rm L})^c = \overline{\psi_{\rm L}} \gamma_\mu \partial^\mu \psi_{\rm L}$. This relationship can easily be proved by taking account of $\partial^\mu \left[\overline{(\psi_{\rm L})^c} \gamma_\mu (\psi_{\rm L})^c\right] = 0$; i.e., we have

$$\begin{aligned} \overline{(\psi_{\rm L})^c} \gamma_\mu \partial^\mu (\psi_{\rm L})^c &= -\partial^\mu \overline{(\psi_{\rm L})^c} \gamma_\mu (\psi_{\rm L})^c = -\left[\partial^\mu \overline{(\psi_{\rm L})^c} \gamma_\mu (\psi_{\rm L})^c\right]^T \\ &= \left(\mathcal{C} \overline{\psi_{\rm L}}^T\right)^T \gamma_\mu^T \partial^\mu \left[(\psi_{\rm L})^T \mathcal{C}\right]^T = \overline{\psi_{\rm L}} \gamma_\mu \partial^\mu \psi_{\rm L} , \end{aligned} \tag{20}$$

where $\mathcal{C}^T \gamma_\mu^T \mathcal{C}^T = \gamma_\mu$, which may be read off from Eq. (4), has been used.

It is worth pointing out that the factor $1/2$ in $\mathcal{L}'_{\rm Majorana}$ allows us to get the Dirac equation of massive Majorana neutrinos analogous to that of massive Dirac neutrinos. To see this point more clearly, let us consider the Lagrangian of free Majorana neutrinos (i.e., their kinetic and mass terms):

$$\begin{aligned} \mathcal{L}_\nu &= i \overline{\nu_{\rm L}} \gamma_\mu \partial^\mu \nu_{\rm L} - \left[\frac{1}{2} \overline{\nu_{\rm L}} M_{\rm L} (\nu_{\rm L})^c + {\rm h.c.}\right] = i \overline{\nu'_{\rm L}} \gamma_\mu \partial^\mu \nu'_{\rm L} - \left[\frac{1}{2} \overline{\nu'_{\rm L}} \widehat{M}_\nu (\nu'_{\rm L})^c + {\rm h.c.}\right] \\ &= \frac{1}{2} \left(i \overline{\nu'} \gamma_\mu \partial^\mu \nu' - \overline{\nu'} \widehat{M}_\nu \nu'\right) = -\frac{1}{2} \left(i \partial^\mu \overline{\nu'} \gamma_\mu \nu' + \overline{\nu'} \widehat{M}_\nu \nu'\right) \; , \end{aligned} \tag{21}$$

where $\partial^\mu (\overline{\nu'} \gamma_\mu \nu') = 0$ has been used. Then we substitute $\mathcal{L}_\nu$ into the Euler-Lagrange equation

$$\partial^\mu \frac{\partial \mathcal{L}_\nu}{\partial \left(\partial^\mu \overline{\nu'}\right)} - \frac{\partial \mathcal{L}_\nu}{\partial \overline{\nu'}} = 0 \tag{22}$$

and obtain the Dirac equation

$$i\gamma_\mu \partial^\mu \nu' - \widehat{M}_\nu \nu' = 0 \; . \tag{23}$$

More explicitly, $i\gamma_\mu \partial^\mu \nu_k - m_k \nu_k = 0$ holds (for $k = 1, 2, 3$). That is why the factor $1/2$ in $\mathcal{L}'_{\rm Majorana}$ makes sense.

The weak charged-current interactions of three neutrinos in Eq. (5) can now be rewritten in terms of their mass eigenstates $\nu'_{\rm L} = V^\dagger \nu_{\rm L}$. In the chosen basis where the flavor and mass eigenstates of three charged leptons are identical, the expression of $\mathcal{L}_{cc}$ for Majorana neutrinos is the same as that given in Eq. (12) for Dirac neutrinos. The unitary matrix V is just the 3×3 Majorana neutrino mixing matrix, which contains two more irremovable CP-violating phases than the 3×3 Dirac neutrino mixing matrix (see section 4 for detailed discussions).

The most salient feature of massive Majorana neutrinos is lepton number violation. Let us make the global phase transformations

$$l(x) \to e^{i\Phi} l(x) \; , \quad \nu'_{\rm L}(x) \to e^{i\Phi} \nu'_{\rm L}(x) \; , \tag{24}$$

where l stands for the column vector of e, μ and τ fields, and Φ is an arbitrary spacetime-independent phase. One can immediately see that the kinetic term $\mathcal{L}_{\rm kinetic}$ and the charged-current interaction term $\mathcal{L}_{\rm cc}$ are invariant under these transformations, but the mass term $\mathcal{L}'_{\rm Majorana}$ is not invariant because of both $\overline{\nu'_{\rm L}} \to e^{-i\Phi} \overline{\nu'_{\rm L}}$ and $(\nu'_{\rm L})^c \to e^{-i\Phi} (\nu'_{\rm L})^c$. The lepton number is therefore violated for massive Majorana neutrinos. Similar to the case of Dirac neutrinos, the lepton flavor violation of Majorana neutrinos is also described by V.

The $0\nu2\beta$ decay $(A, Z) \to (A, Z+2) + 2e^-$ is a clean signature of the Majorana nature of massive neutrinos. This lepton-number-violating process can occur when there exists neutrino-antineutrino mixing induced by the Majorana mass term (i.e., the neutrino mass eigenstates are self-conjugate, $\overline{\nu}_i = \nu_i$). The effective mass of the $0\nu2\beta$ decay is defined as

$$\langle m \rangle_{ee} \equiv \left| \sum_i m_i V_{ei}^2 \right| \; , \tag{25}$$

where m_i comes from the helicity suppression factor m_i/E for the ν_i exchange between two beta decays with E being the energy of the virtual ν_i neutrino. Current experimental data only yield an upper bound $\langle m \rangle_{ee} < 0.23$ eV (or < 0.85 eV as a more conservative bound) at the 2σ level.

1.4 Hybrid neutrino mass terms

Similar to Eq. (14), the right-handed neutrino field $N_{\rm R}$ and its charge-conjugate counterpart $(N_{\rm R})^c$ can also form a Majorana mass term. Hence it is possible to write out the following hybrid neutrino mass terms in terms of $\nu_{\rm L}$, $N_{\rm R}$, $(\nu_{\rm L})^c$ and $(N_{\rm R})^c$ fields:

$$\begin{aligned} -\mathcal{L}'_{\rm hybrid} &= \overline{\nu_{\rm L}} M_{\rm D} N_{\rm R} + \frac{1}{2} \overline{\nu_{\rm L}} M_{\rm L} (\nu_{\rm L})^c + \frac{1}{2} \overline{(N_{\rm R})^c} M_{\rm R} N_{\rm R} + {\rm h.c.} \\ &= \frac{1}{2} \left[\overline{\nu_{\rm L}} \quad \overline{(N_{\rm R})^c} \right] \begin{pmatrix} M_{\rm L} & M_{\rm D} \\ M_{\rm D}^T & M_{\rm R} \end{pmatrix} \begin{bmatrix} (\nu_{\rm L})^c \\ N_{\rm R} \end{bmatrix} + {\rm h.c.} \; , \end{aligned} \tag{26}$$

where $M_{\rm L}$ and $M_{\rm R}$ are symmetric mass matrices because the corresponding mass terms are of the Majorana type, and the relationship

$$\overline{(N_{\rm R})^c} M_{\rm D}^T (\nu_{\rm L})^c = \left[(N_{\rm R})^T \mathcal{C} M_{\rm D}^T \mathcal{C} \overline{\nu_{\rm L}}^T \right]^T = \overline{\nu_{\rm L}} M_{\rm D} N_{\rm R} \tag{27}$$

has been used. The overall 6×6 mass matrix in Eq. (26) is also symmetric, and thus it can be diagonalized by a 6×6 unitary matrix through the transformation

$$\begin{pmatrix} V & R \\ S & U \end{pmatrix}^\dagger \begin{pmatrix} M_{\rm L} & M_{\rm D} \\ M_{\rm D}^T & M_{\rm R} \end{pmatrix} \begin{pmatrix} V & R \\ S & U \end{pmatrix}^* = \begin{pmatrix} \widehat{M}_\nu & \mathbf{0} \\ \mathbf{0} & \widehat{M}_N \end{pmatrix} , \tag{28}$$

where we have defined $\widehat{M}_\nu \equiv {\rm Diag}\{m_1, m_2, m_3\}$, $\widehat{M}_N \equiv {\rm Diag}\{M_1, M_2, M_3\}$, and the 3×3 matrices V, R, S and U satisfy the unitarity conditions

$$\begin{aligned} VV^\dagger + RR^\dagger = SS^\dagger + UU^\dagger &= \mathbf{1} , \\ V^\dagger V + S^\dagger S = R^\dagger R + U^\dagger U &= \mathbf{1} , \\ VS^\dagger + RU^\dagger = V^\dagger R + S^\dagger U &= \mathbf{0} . \end{aligned} \tag{29}$$

After this diagonalization, Eq. (26) becomes

$$-\mathcal{L}'_{\rm hybrid} = \frac{1}{2} \left[\overline{\nu'_{\rm L}} \quad \overline{(N'_{\rm R})^c} \right] \begin{pmatrix} \widehat{M}_\nu & \mathbf{0} \\ \mathbf{0} & \widehat{M}_N \end{pmatrix} \begin{bmatrix} (\nu'_{\rm L})^c \\ N'_{\rm R} \end{bmatrix} + {\rm h.c.} , \tag{30}$$

where $\nu'_{\rm L} = V^\dagger \nu_{\rm L} + S^\dagger (N_{\rm R})^c$ and $N'_{\rm R} = R^T (\nu_{\rm L})^c + U^T N_{\rm R}$ together with $(\nu'_{\rm L})^c = \mathcal{C} \overline{\nu'_{\rm L}}^T$ and $(N'_{\rm R})^c = \mathcal{C} \overline{N'_{\rm R}}^T$. Then the Majorana field

$$\nu' = \begin{bmatrix} \nu'_{\rm L} \\ (N'_{\rm R})^c \end{bmatrix} + \begin{bmatrix} (\nu'_{\rm L})^c \\ N'_{\rm R} \end{bmatrix} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ N_1 \\ N_2 \\ N_3 \end{pmatrix} \tag{31}$$

satisfies the Majorana condition $(\nu')^c = \nu'$ and describes the mass eigenstates of six Majorana neutrinos. In other words,

$$-\mathcal{L}'_{\rm hybrid} = \frac{1}{2} \overline{\nu'} \begin{pmatrix} \widehat{M}_\nu & \mathbf{0} \\ \mathbf{0} & \widehat{M}_N \end{pmatrix} \nu' = \frac{1}{2} \sum_{i=1}^{3} \left(m_i \overline{\nu_i} \nu_i + M_i \overline{N_i} N_i \right) . \tag{32}$$

Because of $\nu_{\rm L} = V \nu'_{\rm L} + R (N'_{\rm R})^c$ and $N_{\rm R} = S^* (\nu'_{\rm L})^c + U^* N'_{\rm R}$, we immediately have $(\nu_{\rm L})^c = V^* (\nu'_{\rm L})^c + R^* N'_{\rm R}$ and $(N_{\rm R})^c = S \nu'_{\rm L} + U (N'_{\rm R})^c$. Given the generic relations $\overline{(\psi_{\rm L})^c} \gamma_\mu \partial^\mu (\psi_{\rm L})^c = \overline{\psi_{\rm L}} \gamma_\mu \partial^\mu \psi_{\rm L}$ and $\overline{(\psi_{\rm R})^c} \gamma_\mu \partial^\mu (\psi_{\rm R})^c = \overline{\psi_{\rm R}} \gamma_\mu \partial^\mu \psi_{\rm R}$ for an arbitrary fermion field ψ, the kinetic term of Majorana neutrinos under consideration turns out to be

$$\begin{aligned} \mathcal{L}_{\rm kinetic} = i \overline{\nu_{\rm L}} \gamma_\mu \partial^\mu \nu_{\rm L} + i \overline{N_{\rm R}} \gamma_\mu \partial^\mu N_{\rm R} = i \overline{\nu'_{\rm L}} \gamma_\mu \partial^\mu \nu'_{\rm L} + i \overline{N'_{\rm R}} \gamma_\mu \partial^\mu N'_{\rm R} = \frac{i}{2} \overline{\nu'} \gamma_\mu \partial^\mu \nu' \\ = \frac{i}{2} \sum_{k=1}^{3} \left(\overline{\nu_k} \gamma_\mu \partial^\mu \nu_k + \overline{N_k} \gamma_\mu \partial^\mu N_k \right) , \end{aligned} \tag{33}$$

where the unitarity conditions given in Eq. (29) have been used.

The weak charged-current interactions of active neutrinos in Eq. (5) can now be rewritten in terms of the mass eigenstates of six Majorana neutrinos via $\nu_{\rm L} = V \nu'_{\rm L} + R (N'_{\rm R})^c$. In the chosen basis where the flavor and mass eigenstates of three charged leptons are identical, we have

$$\mathcal{L}_{\rm cc} = \frac{g}{\sqrt{2}} \overline{(e \ \mu \ \tau)_{\rm L}} \ \gamma^\mu \left[V \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix}_{\rm L} + R \begin{pmatrix} N_1 \\ N_2 \\ N_3 \end{pmatrix}_{\rm L} \right] W^-_\mu + {\rm h.c.} . \tag{34}$$

Note that V and R are responsible for the charged-current interactions of three known neutrinos ν_i and three new neutrinos N_i (for $i = 1, 2, 3$), respectively. Their correlation is described by $VV^\dagger + RR^\dagger = \mathbf{1}$, and thus V is not unitary unless ν_i and N_i are completely decoupled (i.e., $R = \mathbf{0}$).

As a consequence of lepton number violation, the $0\nu2\beta$ decay $(A, Z) \to (A, Z + 2) + 2e^-$ can now take place via the exchanges of both ν_i and N_i between two beta decays, whose coupling matrix elements are V_{ei} and R_{ei} respectively. The relative contributions of ν_i and N_i to this lepton-number-violating process depend not only on m_i, M_i, V_{ei} and R_{ei} but also on the relevant nuclear matrix elements which cannot be reliably evaluated. For a realistic seesaw mechanism working at the TeV scale (i.e., $M_i \sim \mathcal{O}(1)$ TeV) or at a superhigh-energy scale, however, the contribution of ν_i to the $0\nu2\beta$ decay is in most cases dominant.

The hybrid neutrino mass terms in Eq. (26) provide us with the necessary ingredients of a dynamic mechanism to interpret why three known neutrinos have non-zero but tiny masses. The key point is that the mass scales of $M_{\rm L}$, $M_{\rm D}$ and $M_{\rm R}$ may have a strong hierarchy. First, $M_{\rm D} \sim \langle H \rangle \approx 174$ GeV is naturally characterized by the electroweak symmetry breaking scale. Second, $M_{\rm L} \ll \langle H \rangle$ satisfies 't Hooft's naturalness criterion because this Majorana mass term violates lepton number conservation. Third, $M_{\rm R} \gg \langle H \rangle$ is naturally expected since right-handed neutrinos are $SU(2)_{\rm L}$ gauge singlets and thus their mass term is not subject to the electroweak symmetry breaking scale. The hierarchy $M_{\rm R} \gg M_{\rm D} \gg M_{\rm L}$ can therefore allow us to make reliable approximations in deriving the effective mass matrix of three active neutrinos $(\nu_e, \nu_\mu, \nu_\tau)$ from Eq. (28). The latter yields

$$\begin{aligned} R\widehat{M}_N &= M_{\rm L} R^* + M_{\rm D} U^* \; , \\ S\widehat{M}_\nu &= M_{\rm D}^T V^* + M_{\rm R} S^* \; ; \end{aligned} \tag{35}$$

and

$$\begin{aligned} U\widehat{M}_N &= M_{\rm R} U^* + M_{\rm D}^T R^* \; , \\ V\widehat{M}_\nu &= M_{\rm L} V^* + M_{\rm D} S^* \; . \end{aligned} \tag{36}$$

Given $M_{\rm R} \gg M_{\rm D} \gg M_{\rm L}$, $R \sim S \sim \mathcal{O}(M_{\rm D}/M_{\rm R})$ naturally holds, implying that U and V are almost unitary up to the accuracy of $\mathcal{O}(M_{\rm D}^2/M_{\rm R}^2)$. Hence Eq. (36) leads to

$$\begin{aligned} U\widehat{M}_N U^T &= M_{\rm R} (UU^\dagger)^T + M_{\rm D}^T (R^* U^T) \approx M_{\rm R} \; , \\ V\widehat{M}_\nu V^T &= M_{\rm L} (VV^\dagger)^T + M_{\rm D} (S^* V^T) \approx M_{\rm L} + M_{\rm D} (S^* V^T) \; . \end{aligned} \tag{37}$$

$S^* V^T = M_{\rm R}^{-1} S \widehat{M}_\nu V^T - M_{\rm R}^{-1} M_{\rm D}^T (VV^\dagger)^T \approx -M_{\rm R}^{-1} M_{\rm D}^T$ can be derived from Eq. (35). We substitute this expression into Eq. (37) and then obtain

$$M_\nu \equiv V \widehat{M}_\nu V^T \approx M_{\rm L} - M_{\rm D} M_{\rm R}^{-1} M_{\rm D}^T \; . \tag{38}$$

This result, known as the type-(I+II) seesaw relation, is just the effective mass matrix of three light neutrinos. The small mass scale of M_ν is attributed to the small mass scale of $M_{\rm L}$ and the large mass scale of $M_{\rm R}$. There are two particularly interesting limits: (1) If $M_{\rm L}$ is absent from Eq. (26), one will be left with the canonical or type-I seesaw relation $M_\nu \approx -M_{\rm D} M_{\rm R}^{-1} M_{\rm D}^T$; (2) If only $M_{\rm L}$ is present in Eq. (26), one will get the type-II seesaw relation $M_\nu = M_{\rm L}$. More detailed discussions about various seesaw mechanisms and their phenomenological consequences will be presented in sections 6, 7 and 8.

2 Diagnosis of CP Violation

2.1 C, P and T transformations

We begin with a brief summary of the transformation properties of quantum fields under the discrete space-time symmetries of parity (P), charge conjugation (C) and time reversal (T). The parity transformation changes the space coordinates $\vec{x}$ into $-\vec{x}$. The charge conjugation flips the signs of internal

charges of a particle, such as the electric charge and the lepton (baryon) number. The time reversal reflects the time coordinate t into $-t$.

A free Dirac spinor $\psi(t,\vec{x})$ or $\overline{\psi}(t,\vec{x})$ transforms under C, P and T as [3]

$$\begin{aligned}
\psi(t,\vec{x}) &\xrightarrow{\rm C} \mathcal{C}\overline{\psi}^T(t,\vec{x}) \; , \\
\overline{\psi}(t,\vec{x}) &\xrightarrow{\rm C} -\psi^T(t,\vec{x})\mathcal{C}^{-1} \; , \\
\psi(t,\vec{x}) &\xrightarrow{\rm P} \mathcal{P}\psi(t,-\vec{x}) \; , \\
\overline{\psi}(t,\vec{x}) &\xrightarrow{\rm P} \overline{\psi}(t,-\vec{x})\mathcal{P}^\dagger \; , \\
\psi(t,\vec{x}) &\xrightarrow{\rm T} \mathcal{T}\psi(-t,\vec{x}) \; , \\
\overline{\psi}(t,\vec{x}) &\xrightarrow{\rm T} \overline{\psi}(-t,\vec{x})\mathcal{T}^\dagger \; ,
\end{aligned} \tag{39}$$

where $\mathcal{C} = i\gamma_2\gamma_0$, $\mathcal{P} = \gamma_0$ and $\mathcal{T} = \gamma_1\gamma_3$ in the Dirac-Pauli representation. These transformation properties can simply be deduced from the requirement that the Dirac equation $i\gamma_\mu\partial^\mu\psi(t,\vec{x}) = m\psi(t,\vec{x})$ be invariant under C, P or T operation. Note that all the classical numbers (or c-numbers), such as the coupling constants and γ-matrix elements, must be complex-conjugated under T. Note also that the charge-conjugation matrix $\mathcal{C}$ satisfies the conditions given in Eq. (4). It is very important to figure out how the Dirac spinor bilinears transform under C, P and T, because both leptons and quarks are described by spinor fields and they always appear in the bilinear forms in a Lorentz-invariant Lagrangian. Let us consider the following scalar-, pseudoscalar-, vector-, pseudovector- and tensor-like spinor bilinears: $\overline{\psi_1}\psi_2$, $i\overline{\psi_1}\gamma_5\psi_2$, $\overline{\psi_1}\gamma_\mu\psi_2$, $\overline{\psi_1}\gamma_\mu\gamma_5\psi_2$ and $\overline{\psi_1}\sigma_{\mu\nu}\psi_2$, where $\sigma_{\mu\nu} \equiv i[\gamma_\mu,\gamma_\nu]/2$ is defined. One may easily verify that all these bilinears are Hermitian. Under C, P and T, for example,

$$\begin{aligned}
\overline{\psi_1}\gamma_\mu\psi_2 \xrightarrow{\rm C} -\psi_1^T\mathcal{C}^{-1}\gamma_\mu\mathcal{C}\overline{\psi_2}^T = \psi_1^T\gamma_\mu^T\overline{\psi_2}^T = -\left[\overline{\psi_2}\gamma_\mu\psi_1\right]^T = -\overline{\psi_2}\gamma_\mu\psi_1 \; , \\
\overline{\psi_1}\gamma_\mu\psi_2 \xrightarrow{\rm P} \overline{\psi_1}\gamma_0\gamma_\mu\gamma_0\psi_2 = \overline{\psi_1}\gamma^\mu\psi_2 \; , \\
\overline{\psi_1}\gamma_\mu\psi_2 \xrightarrow{\rm T} \overline{\psi_1}\left(\gamma_1\gamma_3\right)^\dagger\gamma_\mu^*\left(\gamma_1\gamma_3\right)\psi_2 = \overline{\psi_1}\gamma^\mu\psi_2 \; ;
\end{aligned} \tag{40}$$

and thus

$$\begin{aligned}
\overline{\psi_1}\gamma_\mu\psi_2 &\xrightarrow{\rm CP} -\overline{\psi_2}\gamma^\mu\psi_1 \; , \\
\overline{\psi_1}\gamma_\mu\psi_2 &\xrightarrow{\rm CPT} -\overline{\psi_2}\gamma_\mu\psi_1 \; ,
\end{aligned} \tag{41}$$

with $\vec{x} \to -\vec{x}$ under P and $t \to -t$ under T for ψ_1 and ψ_2. The transformation properties of five spinor bilinears under C, P, T, CP and CPT are summarized in Table 1, where one should keep in mind that all the c-numbers are complex-conjugated under T and CPT.

It is well known that CPT is a good symmetry in a local quantum field theory which is Lorentz-invariant and possesses a Hermitian Lagrangian. The latter is necessary in order to have a unitary transition operator (i.e., the S-matrix). The CPT invariance of a theory implies that CP and T must be simultaneously conserving or broken, as already examined in the quark sector of the SM via the K^0-$\bar{K}^0$ mixing system. After a slight modification of the SM by introducing the Dirac or Majorana mass term for three neutrinos, one may also look at possible sources of CP or T violation in the lepton sector.

2.2 The source of CP violation

The SM of electroweak interactions is based on the $SU(2)_{\rm L} \times U(1)_{\rm Y}$ gauge symmetry and the Higgs mechanism. The latter triggers the spontaneous symmetry breaking $SU(2)_{\rm L} \times U(1)_{\rm Y} \to U(1)_{\rm Q}$, such

[3] For simplicity, here we have omitted a phase factor associated with each transformation. Because one is always interested in the spinor bilinears, the relevant phase factor usually plays no physical role.

Table 1: Transformation properties of the scalar-, pseudoscalar-, vector-, pseudovector- and tensor-like spinor bilinears under C, P and T. Here $\vec{x} \to -\vec{x}$ under P, CP and CPT, together with $t \to -t$ under T and CPT, is hidden and self-explaining for ψ_1 and ψ_2.

	$\overline{\psi_1}\psi_2$	$i\overline{\psi_1}\gamma_5\psi_2$	$\overline{\psi_1}\gamma_\mu\psi_2$	$\overline{\psi_1}\gamma_\mu\gamma_5\psi_2$	$\overline{\psi_1}\sigma_{\mu\nu}\psi_2$
C	$\overline{\psi_2}\psi_1$	$i\overline{\psi_2}\gamma_5\psi_1$	$-\overline{\psi_2}\gamma_\mu\psi_1$	$\overline{\psi_2}\gamma_\mu\gamma_5\psi_1$	$-\overline{\psi_2}\sigma_{\mu\nu}\psi_1$
P	$\overline{\psi_1}\psi_2$	$-i\overline{\psi_1}\gamma_5\psi_2$	$\overline{\psi_1}\gamma^\mu\psi_2$	$-\overline{\psi_1}\gamma^\mu\gamma_5\psi_2$	$\overline{\psi_1}\sigma^{\mu\nu}\psi_2$
T	$\overline{\psi_1}\psi_2$	$-i\overline{\psi_1}\gamma_5\psi_2$	$\overline{\psi_1}\gamma^\mu\psi_2$	$\overline{\psi_1}\gamma^\mu\gamma_5\psi_2$	$-\overline{\psi_1}\sigma^{\mu\nu}\psi_2$
CP	$\overline{\psi_2}\psi_1$	$-i\overline{\psi_2}\gamma_5\psi_1$	$-\overline{\psi_2}\gamma^\mu\psi_1$	$-\overline{\psi_2}\gamma^\mu\gamma_5\psi_1$	$-\overline{\psi_2}\sigma^{\mu\nu}\psi_1$
CPT	$\overline{\psi_2}\psi_1$	$i\overline{\psi_2}\gamma_5\psi_1$	$-\overline{\psi_2}\gamma_\mu\psi_1$	$-\overline{\psi_2}\gamma_\mu\gamma_5\psi_1$	$\overline{\psi_2}\sigma_{\mu\nu}\psi_1$

that three gauge bosons, three charged leptons and six quarks can all acquire masses. But this mechanism itself does not spontaneously break CP, and thus one may examine the source of CP violation in the SM either before or after spontaneous symmetry breaking.

The Lagrangian of the SM $\mathcal{L} = \mathcal{L}_{\rm G} + \mathcal{L}_{\rm H} + \mathcal{L}_{\rm F} + \mathcal{L}_{\rm Y}$ is composed of four parts: the kinetic term of the gauge fields and their self-interactions ($\mathcal{L}_{\rm G}$), the kinetic term of the Higgs doublet and its potential and interactions with the gauge fields ($\mathcal{L}_{\rm H}$), the kinetic term of the fermion fields and their interactions with the gauge fields ($\mathcal{L}_{\rm F}$), and the Yukawa interactions of the fermion fields with the Higgs doublet ($\mathcal{L}_{\rm Y}$):

$$\begin{aligned}
\mathcal{L}_{\rm G} &= -\frac{1}{4}\left(W^{i\mu\nu}W^i_{\mu\nu} + B^{\mu\nu}B_{\mu\nu}\right) \ , \\
\mathcal{L}_{\rm H} &= \left(D^\mu H\right)^\dagger \left(D_\mu H\right) - \mu^2 H^\dagger H - \lambda\left(H^\dagger H\right)^2 \ , \\
\mathcal{L}_{\rm F} &= \overline{Q_{\rm L}} i\not{D} Q_{\rm L} + \overline{\ell_{\rm L}} i\not{D} \ell_{\rm L} + \overline{U_{\rm R}} i\not{\partial}' U_{\rm R} + \overline{D_{\rm R}} i\not{\partial}' D_{\rm R} + \overline{E_{\rm R}} i\not{\partial}' E_{\rm R} \ , \\
\mathcal{L}_{\rm Y} &= -\overline{Q_{\rm L}} Y_{\rm u} \tilde{H} U_{\rm R} - \overline{Q_{\rm L}} Y_{\rm d} H D_{\rm R} - \overline{\ell_{\rm L}} Y_l H E_{\rm R} + {\rm h.c.} \ ,
\end{aligned} \tag{42}$$

whose notations are self-explanatory. To accommodate massive neutrinos, the simplest way is to slightly modify the $\mathcal{L}_{\rm F}$ and $\mathcal{L}_{\rm Y}$ parts (e.g., by introducing three right-handed neutrinos into the SM and allowing for the Yukawa interactions between neutrinos and the Higgs doublet). CP violation is due to the coexistence of $\mathcal{L}_{\rm F}$ and $\mathcal{L}_{\rm Y}$.

We first show that $\mathcal{L}_{\rm G}$ is always invariant under CP. The transformation properties of gauge fields B_μ and W^i_μ under C and P are

$$\begin{aligned}
\left[B_\mu,\ W^1_\mu,\ W^2_\mu,\ W^3_\mu\right] &\xrightarrow{\rm C} \left[-B_\mu,\ -W^1_\mu,\ +W^2_\mu,\ -W^3_\mu\right] \ , \\
\left[B_\mu,\ W^1_\mu,\ W^2_\mu,\ W^3_\mu\right] &\xrightarrow{\rm P} \left[B^\mu,\ W^{1\mu},\ W^{2\mu},\ W^{3\mu}\right] \ , \\
\left[B_\mu,\ W^1_\mu,\ W^2_\mu,\ W^3_\mu\right] &\xrightarrow{\rm CP} \left[-B^\mu,\ -W^{1\mu},\ +W^{2\mu},\ -W^{3\mu}\right]
\end{aligned} \tag{43}$$

with $\vec{x} \to -\vec{x}$ under P and CP for relevant fields. Then the gauge field tensors $B_{\mu\nu}$ and $W^i_{\mu\nu}$ transform under CP as follows:

$$\left[B_{\mu\nu},\ W^1_{\mu\nu},\ W^2_{\mu\nu},\ W^3_{\mu\nu}\right] \xrightarrow{\rm CP} \left[-B^{\mu\nu},\ -W^{1\mu\nu},\ +W^{2\mu\nu},\ -W^{3\mu\nu}\right] \ . \tag{44}$$

Hence $\mathcal{L}_{\rm G}$ is formally invariant under CP.

We proceed to show that $\mathcal{L}_{\rm H}$ is also invariant under CP. The Higgs doublet H contains two scalar components ϕ^+ and ϕ^0; i.e.,

$$H = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \ , \quad H^\dagger = \begin{pmatrix} \phi^- & \phi^{0*} \end{pmatrix} \ . \tag{45}$$

Therefore,

$$H(t,\vec{x}) \xrightarrow{\rm CP} H^*(t,-\vec{x}) = \begin{pmatrix} \phi^- \\ \phi^{0*} \end{pmatrix} . \tag{46}$$

It is very trivial to prove that the $H^\dagger H$ and $(H^\dagger H)^2$ terms of $\mathcal{L}_{\rm H}$ are CP-invariant. To examine how the $(D^\mu H)^\dagger (D_\mu H)$ term of $\mathcal{L}_{\rm H}$ transforms under CP, we explicitly write out

$$D_\mu H = \left(\partial_\mu - ig\tau^k W^k_\mu - ig' Y B_\mu\right) H = \begin{pmatrix} \partial_\mu \phi^+ - iX^+_\mu \phi^0 - iY^+_\mu \phi^+ \\ \partial_\mu \phi^0 - iX^-_\mu \phi^+ + iY^-_\mu \phi^0 \end{pmatrix} \tag{47}$$

with $X^\pm_\mu \equiv gW^\pm_\mu/\sqrt{2} = g(W^1_\mu \mp iW^2_\mu)/2$, $Y^\pm \equiv \pm g' Y B_\mu + gW^3_\mu/2$, and $k = 1, 2, 3$. Note that

$$X^\pm_\mu \xrightarrow{\rm CP} -X^{\mp\mu} , \quad Y^\pm_\mu \xrightarrow{\rm CP} -Y^{\pm\mu} , \tag{48}$$

together with $\partial_\mu \to \partial^\mu$, $\phi^\pm \to \phi^\mp$ and $\phi^0 \to \phi^{0*}$ under CP. So it is easy to check that $(D^\mu H)^\dagger (D_\mu H)$ is also CP-invariant. Therefore, $\mathcal{L}_{\rm H}$ is formally invariant under CP.

The next step is to examine the CP invariance of $\mathcal{L}_{\rm F}$. To be more specific, we divide $\mathcal{L}_{\rm F}$ into the quark sector and the lepton sector; i.e., $\mathcal{L}_{\rm F} = \mathcal{L}_q + \mathcal{L}_l$. We only analyze the CP property of $\mathcal{L}_q$ in the following, because that of $\mathcal{L}_l$ can be analyzed in the same way. The explicit form of $\mathcal{L}_q$ reads

$$\begin{aligned} \mathcal{L}_q = \overline{Q_{\rm L}} i\not{D} Q_{\rm L} + \overline{U_{\rm R}} i\not{\partial}' U_{\rm R} + \overline{D_{\rm R}} i\not{\partial}' D_{\rm R} = \sum_{j=1}^{3} \Big\{ & \frac{g}{2} \left[\overline{q'_j} \gamma^\mu P_{\rm L} W^1_\mu q_j + \overline{q_j} \gamma^\mu P_{\rm L} W^1_\mu q'_j \right] \\ & + \frac{g}{2} \left[i\overline{q'_j} \gamma^\mu P_{\rm L} W^2_\mu q_j - i\overline{q_j} \gamma^\mu P_{\rm L} W^2_\mu q'_j \right] \\ & + \frac{g}{2} \left[\overline{q_j} \gamma^\mu P_{\rm L} W^3_\mu q_j - \overline{q'_j} \gamma^\mu P_{\rm L} W^3_\mu q'_j \right] \\ & + i \left[\overline{q_j} \gamma^\mu P_{\rm L} \left(\partial_\mu - i\frac{g'}{6} B_\mu \right) q_j \right] \\ & + i \left[\overline{q'_j} \gamma^\mu P_{\rm L} \left(\partial_\mu - i\frac{g'}{6} B_\mu \right) q'_j \right] \\ & + i \left[\overline{q_j} \gamma^\mu P_{\rm R} \left(\partial_\mu - i\frac{2g'}{3} B_\mu \right) q_j \right] \\ & + i \left[\overline{q'_j} \gamma^\mu P_{\rm R} \left(\partial_\mu + i\frac{g'}{3} B_\mu \right) q'_j \right] \Big\} , \end{aligned} \tag{49}$$

where q_j and q'_j (for $j = 1, 2, 3$) run over (u, c, t) and (d, s, b), respectively. The transformation properties of gauge fields B_μ and W^i_μ under C and P have been given in Eq. (43). With the help of Table 1, one can see that the relevant spinor bilinears transform under C and P as follows:

$$\begin{aligned} \overline{\psi_1} \gamma_\mu (1 \pm \gamma_5) \psi_2 & \xrightarrow{\rm C} -\overline{\psi_2} \gamma_\mu (1 \mp \gamma_5) \psi_1 , \\ \overline{\psi_1} \gamma_\mu (1 \pm \gamma_5) \psi_2 & \xrightarrow{\rm P} +\overline{\psi_1} \gamma^\mu (1 \mp \gamma_5) \psi_2 , \\ \overline{\psi_1} \gamma_\mu (1 \pm \gamma_5) \psi_2 & \xrightarrow{\rm CP} -\overline{\psi_2} \gamma^\mu (1 \pm \gamma_5) \psi_1 , \end{aligned} \tag{50}$$

with $\vec{x} \to -\vec{x}$ under P and CP for ψ_1 and ψ_2. Furthermore,

$$\begin{aligned} \overline{\psi_1} \gamma_\mu (1 \pm \gamma_5) \partial^\mu \psi_2 & \xrightarrow{\rm C} \overline{\psi_2} \gamma_\mu (1 \mp \gamma_5) \partial^\mu \psi_1 , \\ \overline{\psi_1} \gamma_\mu (1 \pm \gamma_5) \partial^\mu \psi_2 & \xrightarrow{\rm P} \overline{\psi_1} \gamma^\mu (1 \mp \gamma_5) \partial_\mu \psi_2 , \\ \overline{\psi_1} \gamma_\mu (1 \pm \gamma_5) \partial^\mu \psi_2 & \xrightarrow{\rm CP} \overline{\psi_2} \gamma^\mu (1 \pm \gamma_5) \partial_\mu \psi_1 , \end{aligned} \tag{51}$$

with $\vec{x} \to -\vec{x}$ under P and CP for ψ_1 and ψ_2. It is straightforward to check that $\mathcal{L}_q$ in Eq. (49) is formally invariant under CP. Following the same procedure and using Eqs. (49), (50) and (51), one can easily show that $\mathcal{L}_l = \overline{\ell_{\rm L}} i \not{\partial} \ell_{\rm L} + \overline{E_{\rm R}} i \not{\partial}' E_{\rm R}$ is also CP-invariant. Thus we conclude that $\mathcal{L}_{\rm F}$ is invariant under CP.

The last step is to examine whether $\mathcal{L}_{\rm Y}$ is CP-conserving or not. Explicitly,

$$\begin{aligned} -\mathcal{L}_{\rm Y} &= \overline{Q_{\rm L}} Y_{\rm u} \tilde{H} U_{\rm R} + \overline{Q_{\rm L}} Y_{\rm d} H D_{\rm R} + \overline{\ell_{\rm L}} Y_l H E_{\rm R} + {\rm h.c.} \\ &= \sum_{j,k=1}^{3} \left\{ (Y_{\rm u})_{jk} \left[\overline{q_j} P_{\rm R} q_k \phi^{0*} - \overline{q'_j} P_{\rm R} q_k \phi^- \right] \right. \\ &\quad + (Y_{\rm u})^*_{jk} \left[\overline{q_k} P_{\rm L} q_j \phi^0 - \overline{q_k} P_{\rm L} q'_j \phi^+ \right] \\ &\quad + (Y_{\rm d})_{jk} \left[\overline{q_j} P_{\rm R} q'_k \phi^+ + \overline{q'_j} P_{\rm R} q'_k \phi^0 \right] \\ &\quad + (Y_{\rm d})^*_{jk} \left[\overline{q'_k} P_{\rm L} q_j \phi^- + \overline{q'_k} P_{\rm L} q'_j \phi^{0*} \right] \\ &\quad + (Y_l)_{jk} \left[\overline{\nu_j} P_{\rm R} l_k \phi^+ + \overline{l_j} P_{\rm R} l_k \phi^0 \right] \\ &\quad \left. + (Y_l)^*_{jk} \left[\overline{l_k} P_{\rm L} \nu_j \phi^- + \overline{l_k} P_{\rm L} l_j \phi^{0*} \right] \right\} , \end{aligned} \tag{52}$$

where q_j and q'_j (for $j = 1, 2, 3$) run over (u, c, t) and (d, s, b), respectively; while ν_j and l_j (for $j = 1, 2, 3$) run over $(\nu_e, \nu_\mu, \nu_\tau)$ and (e, μ, τ), respectively. Because of $\phi^\pm \to \phi^\mp$, $\phi^0 \to \phi^{0*}$ and $\overline{\psi_1}(1 \pm \gamma_5)\psi_2 \to \overline{\psi_2}(1 \mp \gamma_5)\psi_1$ under CP, we immediately arrive at

$$\begin{aligned} -\mathcal{L}_{\rm Y} &\xrightarrow{\rm CP} \sum_{j,k=1}^{3} \left\{ (Y_{\rm u})_{jk} \left[\overline{q_k} P_{\rm L} q_j \phi^0 - \overline{q_k} P_{\rm L} q'_j \phi^+ \right] \right. \\ &\quad + (Y_{\rm u})^*_{jk} \left[\overline{q_j} P_{\rm R} q_k \phi^{0*} - \overline{q'_j} P_{\rm R} q_k \phi^- \right] \\ &\quad + (Y_{\rm d})_{jk} \left[\overline{q'_k} P_{\rm L} q_j \phi^- + \overline{q'_k} P_{\rm L} q'_j \phi^{0*} \right] \\ &\quad + (Y_{\rm d})^*_{jk} \left[\overline{q_j} P_{\rm R} q'_k \phi^+ + \overline{q'_j} P_{\rm R} q'_k \phi^0 \right] \\ &\quad + (Y_l)_{jk} \left[\overline{l_k} P_{\rm L} \nu_j \phi^- + \overline{l_k} P_{\rm L} l_j \phi^{0*} \right] \\ &\quad \left. + (Y_l)^*_{jk} \left[\overline{\nu_j} P_{\rm R} l_k \phi^+ + \overline{l_j} P_{\rm R} l_k \phi^0 \right] \right\} , \end{aligned} \tag{53}$$

with $\vec{x} \to -\vec{x}$ for both scalar and spinor fields under consideration. Comparing between Eqs. (52) and (53), we see that $\mathcal{L}_{\rm Y}$ will be formally invariant under CP if the conditions

$$(Y_{\rm u})_{jk} = (Y_{\rm u})^*_{jk} , \quad (Y_{\rm d})_{jk} = (Y_{\rm d})^*_{jk} , \quad (Y_l)_{jk} = (Y_l)^*_{jk} \tag{54}$$

are satisfied. In other words, the Yukawa coupling matrices $Y_{\rm u}$, $Y_{\rm d}$ and Y_l must be real to guarantee the CP invartiance of $\mathcal{L}_{\rm Y}$. Given three massless neutrinos in the SM, it is always possible to make Y_l real by redefining the phases of charged-lepton fields. But it is in general impossible to make both $Y_{\rm u}$ and $Y_{\rm d}$ real for three families of quarks, and thus CP violation can only appear in the quark sector.

Given massive neutrinos beyond the SM, $\mathcal{L}_{\rm Y}$ must be modified. The simplest way is to introduce three right-handed neutrinos and incorporate the Dirac neutrino mass term in Eq. (6) into $\mathcal{L}_{\rm Y}$. In this case one should also add the kinetic term of three right-handed neutrinos into $\mathcal{L}_{\rm F}$. It is straightforward to show that the conditions of CP invariance in the lepton sector turn out to be

$$Y_\nu = Y^*_\nu , \qquad Y_l = Y^*_l , \tag{55}$$

exactly in parallel with the quark sector. If an effective Majorana mass term is introduced into $\mathcal{L}_{\rm Y}$, as shown in Eq. (14), then the conditions of CP invariance in the lepton sector become

$$M_{\rm L} = M_{\rm L}^* \; , \quad Y_l = Y_l^* \; , \tag{56}$$

where $M_{\rm L}$ is the effective Majorana neutrino mass matrix. One may diagonalize both Y_ν (or $M_{\rm L}$) and Y_l to make them real and positive, but such a treatment will transfer CP violation from the Yukawa interactions to the weak charged-current interactions. Then lepton flavor mixing and CP violation are described by the 3×3 unitary matrix V given in Eq. (12), analogous to the 3×3 unitary matrix of quark flavor mixing and CP violation. In other words, the source of CP violation is the irremovable complex phase(s) in the flavor mixing matrix of quarks or leptons. That is why we claim that CP violation stems from the coexistence of $\mathcal{L}_{\rm F}$ and $\mathcal{L}_{\rm Y}$ within the SM and, in most cases, beyond the SM.

It is worth reiterating that the process of spontaneous gauge symmetry breaking in the SM does not spontaneously violate CP. After the Higgs doublet H acquires its vacuum expectation value (i.e., $\phi^+ \to 0$ and $\phi^0 \to v/\sqrt{2}$ with v being real), we obtain three massive gauge bosons $W_\mu^\pm$ and Z_μ as well as one massless gauge boson A_μ. According to their relations with W_μ^i and B_μ, it is easy to find out the transformation properties of these physical fields under CP:

$$W_\mu^\pm \stackrel{\rm CP}{\longrightarrow} -W^{\mp\mu} \; , \quad Z_\mu \stackrel{\rm CP}{\longrightarrow} -Z^\mu \; , \quad A_\mu \stackrel{\rm CP}{\longrightarrow} -A^\mu \; , \tag{57}$$

with $\vec{x} \to -\vec{x}$ under P and CP for each field. In contrast, the neutral Higgs boson h is a CP-even particle. After spontaneous electroweak symmetry breaking, we are left with the quark mass matrices $M_{\rm u} = v Y_{\rm u}/\sqrt{2}$ and $M_{\rm d} = v Y_{\rm d}/\sqrt{2}$ or the lepton mass matrices $M_{\rm D} = v Y_\nu/\sqrt{2}$ and $M_l = v Y_l/\sqrt{2}$. The conditions of CP invariance given above can therefore be replaced with the corresponding mass matrices.

3 Electromagnetic Properties

3.1 Electromagnetic form factors

Although a neutrino does not possess any electric charge, it can have electromagnetic interactions via quantum loops. One may summarize such interactions by means of the following effective interaction term:

$$\mathcal{L}_{\rm EM} = \overline{\psi} \Gamma_\mu \psi A^\mu \equiv J_\mu(x) A^\mu(x) \; , \tag{58}$$

where the form of the electromagnetic current $J_\mu(x)$ is our present concern. Dirac and Majorana neutrinos couple to the photon in different ways, which are described by their respective electromagnetic form factors.

For an arbitrary Dirac particle (e.g., a Dirac neutrino), let us write down the matrix element of $J_\mu(x)$ between two one-particle states:

$$\langle \psi(p') | J_\mu(x) | \psi(p) \rangle = e^{-iqx} \langle \psi(p') | J_\mu(0) | \psi(p) \rangle = e^{-iqx} \overline{u}(\vec{p}') \Gamma_\mu(p, p') u(\vec{p}) \tag{59}$$

with $q = p - p'$. Because $J_\mu(x)$ is a Lorentz vector, the electromagnetic vertex function $\Gamma_\mu(p, p')$ must be a Lorentz vector too. The electromagnetic current conservation (or $U(1)_{\rm Q}$ gauge symmetry) requires $\partial^\mu J_\mu(x) = 0$, leading to

$$\langle \psi(p') | \partial^\mu J_\mu(x) | \psi(p) \rangle = (-iq^\mu)\, e^{-iqx} \overline{u}(\vec{p}') \Gamma_\mu(p, p') u(\vec{p}) = 0 \; . \tag{60}$$

Thus

$$q^\mu \overline{u}(\vec{p}') \Gamma_\mu(p, p') u(\vec{p}) = 0 \tag{61}$$

holds as one of the model-independent constraints on the form of $\Gamma_\mu(p,p')$. In addition, the Hermiticity of $J_\mu(x)$ or its matrix element implies

$$\begin{aligned} & e^{-iqx}\overline{u}(\vec{p}')\Gamma_\mu(p,p')u(\vec{p}) = e^{+iqx}\left[\overline{u}(\vec{p}')\Gamma_\mu(p,p')u(\vec{p})\right]^\dagger \\ & = e^{+iqx}\overline{u}(\vec{p})\left[\gamma_0\Gamma^\dagger_\mu(p,p')\gamma_0\right]u(\vec{p}') = e^{-iqx}\overline{u}(\vec{p}')\left[\gamma_0\Gamma^\dagger_\mu(p',p)\gamma_0\right]u(\vec{p}) \; , \end{aligned} \tag{62}$$

from which we immediately arrive at the second constraint on $\Gamma_\mu(p,p')$:

$$\Gamma_\mu(p,p') = \gamma_0\Gamma^\dagger_\mu(p',p)\gamma_0 \; . \tag{63}$$

Because of $p^2 = p'^2 = m^2$ with m being the fermion mass, we have $(p+p')^2 = 4m^2 - q^2$. Hence $\Gamma_\mu(p,p')$ depends only on the Lorentz-invariant quantity q^2.

A careful analysis of the Lorentz structure of $\overline{u}(\vec{p}')\Gamma_\mu(p,p')u(\vec{p})$, with the help of the Gordon-like identities and the constraints given above, shows that $\Gamma_\mu(p,p')$ may in general consist of four independent terms:

$$\Gamma_\mu(p,p') = f_{\rm Q}(q^2)\gamma_\mu + f_{\rm M}(q^2)i\sigma_{\mu\nu}q^\nu + f_{\rm E}(q^2)\sigma_{\mu\nu}q^\nu\gamma_5 + f_{\rm A}(q^2)\left(q^2\gamma_\mu - q_\mu \not{q}\right)\gamma_5 \; , \tag{64}$$

where $f_{\rm Q}(q^2)$, $f_{\rm M}(q^2)$, $f_{\rm E}(q^2)$ and $f_{\rm A}(q^2)$ are usually referred to as the charge, magnetic dipole, electric dipole and anapole form factors, respectively. In the non-relativistic limit of $\mathcal{L}_{\rm EM}$, it is easy to find that $f_{\rm Q}(0) = Q$ represents the electric charge of the particle, $f_{\rm M}(0) \equiv \mu$ denotes the magnetic dipole moment of the particle (i.e., $\mathcal{L}_{\rm EM}(f_{\rm M}) = -\mu\vec{\sigma}\cdot\vec{B}$ with $\vec{B}$ being the static magnetic field), $f_{\rm E}(0) \equiv \epsilon$ stands for the electric dipole moment of the particle (i.e., $\mathcal{L}_{\rm EM}(f_{\rm E}) = -\epsilon\vec{\sigma}\cdot\vec{E}$ with $\vec{E}$ being the static electric field), and $f_{\rm A}(0)$ corresponds to the Zeldovich anapole moment of the particle (i.e., $\mathcal{L}_{\rm EM}(f_{\rm A}) \propto f_{\rm A}(0)\vec{\sigma}\cdot[\nabla\times\vec{B} - \dot{\vec{E}}]$). One can observe that these form factors are not only Lorentz-invariant but also real (i.e., ${\rm Im}f_{\rm Q} = {\rm Im}f_{\rm M} = {\rm Im}f_{\rm E} = {\rm Im}f_{\rm A} = 0$). The latter is actually guaranteed by the Hermiticity condition in Eq. (62).

Given the form of Γ_μ in Eq. (64), it is straightforward to check the CP properties of $\mathcal{L}_{\rm EM}$ in Eq. (58). Note that the photon field transforms as $A^\mu \to -A_\mu$ under CP, and [4]

$$\begin{aligned} \overline{\psi}\gamma_\mu\psi & \xrightarrow{\rm CP} -\overline{\psi}\gamma^\mu\psi \; , \\ \overline{\psi}\gamma_\mu\gamma_5\psi & \xrightarrow{\rm CP} -\overline{\psi}\gamma^\mu\gamma_5\psi \; , \\ \overline{\psi}\sigma_{\mu\nu}\psi & \xrightarrow{\rm CP} -\overline{\psi}\sigma^{\mu\nu}\psi \; , \\ \overline{\psi}\sigma_{\mu\nu}\gamma_5\psi & \xrightarrow{\rm CP} +\overline{\psi}\sigma^{\mu\nu}\gamma_5\psi \; . \end{aligned} \tag{65}$$

Hence only the term proportional to $f_{\rm E}$ in $\mathcal{L}_{\rm EM}$ is CP-violating. If CP were conserved, then this term would vanish (i.e., $f_{\rm E} = 0$ would hold). Although there is no experimental hint at CP violation in the lepton sector, we expect that it should exist as in the quark sector. In any case, all four form factors are finite for a Dirac neutrino.

If neutrinos are massive Majorana particles, their electromagnetic properties will be rather different. The reason is simply that Majorana particles are their own antiparticles and thus can be described by using a smaller number of degrees of freedom. A free Majorana neutrino field ψ is by definition equal to its charge-conjugate field $\psi^c = C\overline{\psi}^T$ up to a global phase. Then

$$\overline{\psi}\Gamma_\mu\psi = \overline{\psi^c}\Gamma_\mu\psi^c = \psi^T\mathcal{C}\Gamma_\mu\mathcal{C}\overline{\psi}^T = \left(\psi^T\mathcal{C}\Gamma_\mu\mathcal{C}\overline{\psi}^T\right)^T = -\overline{\psi}\mathcal{C}^T\Gamma^T_\mu\mathcal{C}^T\psi \; , \tag{66}$$

[4] Taking account of $\mathcal{C}^{-1}\sigma_{\mu\nu}\mathcal{C} = -\sigma^T_{\mu\nu}$ and $\mathcal{C}^{-1}\gamma_5\mathcal{C} = \gamma_5^T$, one may easily prove that $\overline{\psi}\sigma_{\mu\nu}\gamma_5\psi$ is odd under both C and P. Thus $\overline{\psi}\sigma_{\mu\nu}\gamma_5\psi$ is CP-even.

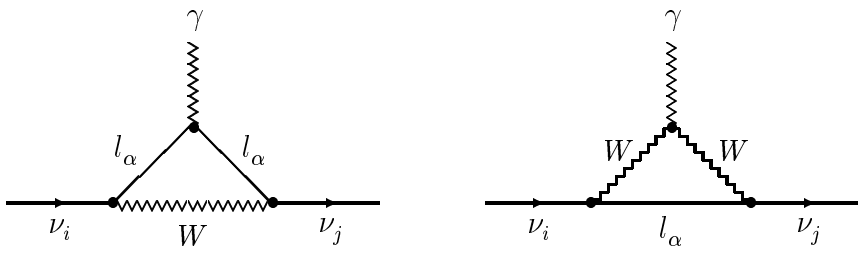


Fig. 1: One-loop Feynman diagrams contributing to the magnetic and electric dipole moments of massive Dirac neutrinos, where $\alpha = e, \mu, \tau$ and $i, j = 1, 2, 3$.

from which one arrives at

$$\Gamma_\mu = -\mathcal{C}^T \Gamma_\mu^T \mathcal{C}^T = \mathcal{C} \Gamma_\mu^T \mathcal{C}^{-1} \; . \tag{67}$$

Substituting Eq. (64) into the right-hand side of Eq. (67) and taking account of $\mathcal{C}\gamma_\mu^T\mathcal{C}^{-1} = -\gamma_\mu$, $\mathcal{C}(\gamma_\mu\gamma_5)^T\mathcal{C}^{-1} = +\gamma_\mu\gamma_5$, $\mathcal{C}\sigma_{\mu\nu}^T\mathcal{C}^{-1} = -\sigma_{\mu\nu}$ and $\mathcal{C}(\sigma_{\mu\nu}\gamma_5)^T\mathcal{C}^{-1} = -\sigma_{\mu\nu}\gamma_5$, we obtain

$$\Gamma_\mu(p, p') = -f_{\rm Q}(q^2)\gamma_\mu - f_{\rm M}(q^2) i\sigma_{\mu\nu} q^\nu - f_{\rm E}(q^2)\sigma_{\mu\nu} q^\nu \gamma_5 + f_{\rm A}(q^2) \left(q^2\gamma_\mu - q_\mu \not{q}\right) \gamma_5 \; . \tag{68}$$

A comparison between Eqs. (64) and (68) yields

$$f_{\rm Q}(q^2) = f_{\rm M}(q^2) = f_{\rm E}(q^2) = 0 \; . \tag{69}$$

This result means that a Majorana neutrino only has the anapole form factor $f_{\rm A}(q^2)$.

More generally, one may write out the matrix elements of the electromagnetic current $J_\mu(x)$ between two different states (i.e., the incoming and outgoing particles are different):

$$\langle \psi_j(p') | J_\mu(x) | \psi_i(p) \rangle = e^{-iqx} \overline{u_j}(\vec{p}\,') \Gamma_\mu^{ij}(p, p') u_i(\vec{p}) \; , \tag{70}$$

where $q = p - p'$ together with $p^2 = m_i^2$ and $p'^2 = m_j^2$ (for $i \neq j$). Here the electromagnetic vertex matrix $\Gamma_\mu(p, p')$ can be decomposed into the following Lorentz-invariant form in terms of four form factors:

$$\Gamma_\mu(p, p') = F_{\rm Q}(q^2) \left(q^2\gamma_\mu - q_\mu \not{q}\right) + F_{\rm M}(q^2) i\sigma_{\mu\nu} q^\nu + F_{\rm E}(q^2)\sigma_{\mu\nu} q^\nu \gamma_5 + F_{\rm A}(q^2) \left(q^2\gamma_\mu - q_\mu \not{q}\right) \gamma_5 \; , \tag{71}$$

where $F_{\rm Q}$, $F_{\rm M}$, $F_{\rm E}$ and $F_{\rm A}$ are all the 2×2 matrices in the space of neutrino mass eigenstates. The diagonal case (i.e., $i = j$) has been discussed above, from Eq. (59) to Eq. (69). In the off-diagonal case (i.e., $i \neq j$), the Hermiticity of $J_\mu(x)$ is no more a constraint on $\Gamma_\mu(p, p')$ for Dirac neutrinos because Eq. (62) only holds for $i = j$. It is now possible for Majorana neutrinos to have finite *transition* dipole moments, simply because Eqs. (66)—(69) do not hold when ψ_i and ψ_j represent different flavors.

We conclude that Dirac neutrinos may have both electric and magnetic dipole moments, while Majorana neutrinos have neither electric nor magnetic dipole moments. But massive Majorana neutrinos can have *transition* dipole moments which involve two different neutrino flavors in the initial and final states, so can massive Dirac neutrinos.

3.2 Magnetic and electric dipole moments

The magnetic and electric dipole moments of massive neutrinos, denoted as $\mu \equiv F_{\rm M}(0)$ and $\epsilon \equiv F_{\rm E}(0)$, are interesting in both theories and experiments because they are closely related to the dynamics of neutrino mass generation and to the characteristic of new physics.

Let us consider a minimal extension of the SM in which three right-handed neutrinos are introduced and lepton number conservation is required. In this case massive neutrinos are Dirac particles and

their magnetic and electric dipole moments can be evaluated by calculating the Feynman diagrams in Fig. 1. Taking account of the smallness of both m^2_α/M^2_W and m^2_i/M^2_W, where m_α (for $\alpha = e, \mu, \tau$) and m_i (for $i = 1, 2, 3$) stand respectively for the charged-lepton and neutrino masses, one obtains

$$\mu^{\rm D}_{ij} = \frac{3eG_{\rm F}m_i}{32\sqrt{2}\pi^2}\left(1+\frac{m_j}{m_i}\right)\times\sum_\alpha\left(2-\frac{m^2_\alpha}{M^2_W}\right)V_{\alpha i}V^*_{\alpha j}\ ,$$
$$\epsilon^{\rm D}_{ij} = \frac{3eG_{\rm F}m_i}{32\sqrt{2}\pi^2}\left(1-\frac{m_j}{m_i}\right)\times\sum_\alpha\left(2-\frac{m^2_\alpha}{M^2_W}\right)V_{\alpha i}V^*_{\alpha j}\ , \tag{72}$$

to an excellent degree of accuracy. Here $V_{\alpha i}$ and $V_{\alpha j}$ are the elements of the unitary lepton flavor mixing matrix V. Some discussions are in order.

(1) In the diagonal case (i.e., $i = j$), we are left with vanishing electric dipole moments (i.e., $\epsilon^{\rm D}_{ii} = 0$). The magnetic dipole moments $\mu^{\rm D}_{ii}$ are finite and proportional to the neutrino masses m_i (for $i = 1, 2, 3$):

$$\mu^{\rm D}_{ii} = \frac{3eG_{\rm F}m_i}{8\sqrt{2}\pi^2}\left(1-\frac{1}{2}\sum_\alpha\frac{m^2_\alpha}{M^2_W}|V_{\alpha i}|^2\right)\ . \tag{73}$$

Hence a massless Dirac neutrino in the SM has no magnetic dipole moment. In the leading-order approximation, $\mu^{\rm D}_{ii}$ are independent of the strength of lepton flavor mixing and have tiny values

$$\mu^{\rm D}_{ii} \approx \frac{3eG_{\rm F}m_i}{8\sqrt{2}\pi^2} \approx 3\times10^{-19}\left(\frac{m_i}{1\ {\rm eV}}\right)\mu_{\rm B}\ , \tag{74}$$

where $\mu_{\rm B} = e\hbar/(2m_e)$ is the Bohr magneton. Given $m_i \leq 1$ eV, the magnitude of $\mu^{\rm D}_{ii}$ is far below its present experimental upper bound ($<$ a few $\times\, 10^{-11}\mu_{\rm B}$).

(2) In the off-diagonal case (i.e., $i \neq j$), the unitarity of V allows us to simplify Eq. (72) to

$$\mu^{\rm D}_{ij} = -\frac{3eG_{\rm F}m_i}{32\sqrt{2}\pi^2}\left(1+\frac{m_j}{m_i}\right)\sum_\alpha\frac{m^2_\alpha}{M^2_W}V_{\alpha i}V^*_{\alpha j}\ ,$$
$$\epsilon^{\rm D}_{ij} = -\frac{3eG_{\rm F}m_i}{32\sqrt{2}\pi^2}\left(1-\frac{m_j}{m_i}\right)\sum_\alpha\frac{m^2_\alpha}{M^2_W}V_{\alpha i}V^*_{\alpha j}\ . \tag{75}$$

We see that the magnitudes of $\mu^{\rm D}_{ij}$ and $\epsilon^{\rm D}_{ij}$ (for $i \neq j$), compared with that of $\mu^{\rm D}_{ii}$, are further suppressed due to the smallness of m^2_α/M^2_W. Similar to the expression given in Eq. (74),

$$\mu^{\rm D}_{ij} \approx -4\times10^{-23}\left(\frac{m_i+m_j}{1\ {\rm eV}}\right)\times\left(\sum_\alpha\frac{m^2_\alpha}{m^2_\tau}V_{\alpha i}V^*_{\alpha j}\right)\mu_{\rm B}\ ,$$
$$\epsilon^{\rm D}_{ij} \approx -4\times10^{-23}\left(\frac{m_i-m_j}{1\ {\rm eV}}\right)\times\left(\sum_\alpha\frac{m^2_\alpha}{m^2_\tau}V_{\alpha i}V^*_{\alpha j}\right)\mu_{\rm B}\ , \tag{76}$$

which can illustrate how small $\mu^{\rm D}_{ij}$ and $\epsilon^{\rm D}_{ij}$ are.

(3) Although Majorana neutrinos do not have intrinsic ($i = j$) magnetic and electric dipole moments, they may have finite transition ($i \neq j$) dipole moments. Because of the fact that Majorana neutrinos are their own antiparticles, their magnetic and electric dipole moments can also get contributions from two additional one-loop Feynman diagrams involving the charge-conjugate fields of ν_i, ν_j, l_α, $W^\pm$ and γ shown in Fig. 1 [5]. In this case one obtains

$$\mu^{\rm M}_{ij} = -\frac{3eG_{\rm F}i}{16\sqrt{2}\pi^2}\left(m_i+m_j\right)\times\sum_\alpha\frac{m^2_\alpha}{M^2_W}{\rm Im}\left(V_{\alpha i}V^*_{\alpha j}\right)\ ,$$

[5] Here we confine ourselves to a simple extension of the SM with three known neutrinos to be massive Majorana particles.

$$\epsilon^{\rm M}_{ij} = -\frac{3eG_{\rm F}}{16\sqrt{2}\pi^2}\left(m_i - m_j\right) \times \sum_\alpha \frac{m^2_\alpha}{M^2_W} {\rm Re}\left(V_{\alpha i} V^*_{\alpha j}\right) \; , \tag{77}$$

where $m_i \neq m_j$ must hold. Comparing between Eqs. (75) and (77), we observe that the magnitudes of $\mu^{\rm M}_{ij}$ and $\epsilon^{\rm M}_{ij}$ are the same order as those of $\mu^{\rm D}_{ij}$ and $\epsilon^{\rm D}_{ij}$ in most cases, although the CP-violating phases hidden in $V_{\alpha i} V^*_{\alpha j}$ are possible to give rise to significant cancellations in some cases.

(4) The fact that μ_{ij} and ϵ_{ij} are proportional to m_i or m_j can be understood in the following way. Note that both tensor- and pseudotensor-like spinor bilinears are chirality-changing operators, which link the left-handed state to the right-handed one [6]:

$$\begin{aligned} \overline{\psi} \sigma_{\mu\nu} \psi &= \overline{\psi_{\rm L}} \sigma_{\mu\nu} \psi_{\rm R} + {\rm h.c.} \; , \\ \overline{\psi} \sigma_{\mu\nu} \gamma_5 \psi &= \overline{\psi_{\rm L}} \sigma_{\mu\nu} \gamma_5 \psi_{\rm R} - {\rm h.c.} \; . \end{aligned} \tag{78}$$

Note also that the same relations hold when ψ is replaced by its charge-conjugate field ψ^c for Majorana neutrinos. Because $(\nu_i)_{\rm R}$ and $(\nu_j)_{\rm R}$ do not have any interactions with $W^\pm$ in Fig. 1, it seems that only $(\nu_i)_{\rm L}$ and $(\nu_j)_{\rm L}$ are flowing along the external fermion lines. To obtain a chirality-changing contribution from the effective (one-loop) electromagnetic vertex, one has to put a mass insertion on one of the external legs in the Feynman diagrams. As a result, the magnetic and electric dipole moments must involve m_i and m_j, the masses of ν_i and ν_j neutrinos.

(5) Is the magnetic or electric dipole moment of a neutrino always proportional to its mass? The answer is negative if new physics beyond the $SU(2)_{\rm L} \times U(1)_{\rm Y}$ gauge theory is involved. For instance, a new term proportional to the charged-lepton mass can contribute to the magnetic dipole moment of a massive Dirac neutrino in the $SU(2)_{\rm L} \times SU(2)_{\rm R} \times U(1)_{\rm Y}$ model with broken left-right symmetry. Depending on the details of this model, such a term might cancel or exceed the term proportional to the neutrino mass in the expression of the magnetic dipole moment.

Finite magnetic and electric dipole moments of massive neutrinos may produce a variety of new processes beyond the SM. For example, (a) radiative neutrino decays $\nu_i \to \nu_j + \gamma$ can happen, so can the Cherenkov radiation of neutrinos in an external electromagnetic field; (b) the elastic neutrino-electron or neutrino-nucleon scattering can be mediated by the magnetic and electric dipole moments; (c) the phenomenon of precession of the neutrino spin can occur in an external magnetic field; (d) the photon (or plasmon) can decay into a neutrino-antineutrino pair in a plasma (i.e., $\gamma^* \to \nu\overline{\nu}$). Of course, non-vanishing electromagnetic dipole moments contribute to neutrino masses too.

3.3 Radiative neutrino decays

If the electromagnetic moments of a massive neutrino ν_i are finite, it can decay into a lighter neutrino ν_j and a photon γ. The Lorentz-invariant vertex matrix of this $\nu_i \to \nu_j + \gamma$ process is in general described by $\Gamma_\mu(p, p')$ in Eq. (71). Because $q^2 = 0$ and $q_\mu \varepsilon^\mu = 0$ hold for a real photon γ, where ε^μ represents the photon polarization, the form of $\Gamma_\mu(p, p')$ can be simplified to

$$\Gamma_\mu(p, p') = \left[i F_{\rm M}(0) + F_{\rm E}(0) \gamma_5 \right] \sigma_{\mu\nu} q^\nu \; . \tag{79}$$

By definition, $F^{ij}_{\rm M}(0) \equiv \mu_{ij}$ and $F^{ij}_{\rm E}(0) \equiv \epsilon_{ij}$ are just the magnetic and electric transition dipole moments between ν_i and ν_j neutrinos. Given the transition matrix element $\overline{u_j}(\vec{p}\,') \Gamma^{ij}_\mu(p, p') u_i(\vec{p})$, it is straightforward to calculate the decay rate. In the rest frame of the decaying neutrino ν_i,

$$\Gamma_{\nu_i \to \nu_j + \gamma} = \frac{\left(m^2_i - m^2_j\right)^3}{8\pi m^3_i} \left(\left|\mu_{ij}\right|^2 + \left|\epsilon_{ij}\right|^2 \right) \; . \tag{80}$$

[6] That is why both magnetic and electric dipole moments must vanish for a Weyl neutrino, because it is massless and does not possess the right-handed component.

This result is valid for both Dirac and Majorana neutrinos.

In the $SU(2)_{\rm L} \times U(1)_{\rm Y}$ gauge theory with three massive Dirac (or Majorana) neutrinos, the radiative decay $\nu_i \to \nu_j + \gamma$ is mediated by the one-loop Feynman diagrams (and their charge-conjugate diagrams) shown in Fig. 1. The explicit expressions of μ_{ij} and ϵ_{ij} have been given in Eq. (75) for Dirac neutrinos and in Eq. (77) for Majorana neutrinos. Hence

$$\Gamma^{\rm (D)}_{\nu_i \to \nu_j + \gamma} = \frac{\left(m_i^2 - m_j^2\right)^3}{8\pi m_i^3} \left(\left|\mu^{\rm D}_{ij}\right|^2 + \left|\epsilon^{\rm D}_{ij}\right|^2\right) = \frac{9\alpha G_{\rm F}^2 m_i^5}{2^{11}\pi^4} \left(1 - \frac{m_j^2}{m_i^2}\right)^3 \left(1 + \frac{m_j^2}{m_i^2}\right) \times \left| \sum_\alpha \frac{m_\alpha^2}{M_W^2} V_{\alpha i} V^*_{\alpha j} \right|^2 , \tag{81}$$

for Dirac neutrinos; or

$$\Gamma^{\rm (M)}_{\nu_i \to \nu_j + \gamma} = \frac{\left(m_i^2 - m_j^2\right)^3}{8\pi m_i^3} \left(\left|\mu^{\rm M}_{ij}\right|^2 + \left|\epsilon^{\rm M}_{ij}\right|^2\right) = \frac{9\alpha G_{\rm F}^2 m_i^5}{2^{10}\pi^4} \left(1 - \frac{m_j^2}{m_i^2}\right)^3 \left\{ \left(1 + \frac{m_j}{m_i}\right)^2 \times \left[\sum_\alpha \frac{m_\alpha^2}{M_W^2} {\rm Im}\left(V_{\alpha i} V^*_{\alpha j}\right) \right]^2 + \left(1 - \frac{m_j}{m_i}\right)^2 \left[\sum_\alpha \frac{m_\alpha^2}{M_W^2} {\rm Re}\left(V_{\alpha i} V^*_{\alpha j}\right) \right]^2 \right\} , \tag{82}$$

for Majorana neutrinos, where $\alpha = e^2/(4\pi)$ denotes the electromagnetic fine-structure constant.

To compare $\Gamma_{\nu_i \to \nu_j + \gamma}$ with the experimental data in a simpler way, one may define an effective magnetic dipole moment

$$\mu_{\rm eff} \equiv \sqrt{\left|\mu_{ij}\right|^2 + \left|\epsilon_{ij}\right|^2} \; . \tag{83}$$

Eq. (80) can then be expressed as

$$\Gamma_{\nu_i \to \nu_j + \gamma} = 5.3 \times \left(1 - \frac{m_j^2}{m_i^2}\right)^3 \left(\frac{m_i}{1~{\rm eV}}\right)^3 \times \left(\frac{\mu_{\rm eff}}{\mu_{\rm B}}\right)^2 {\rm s}^{-1} \; . \tag{84}$$

Although $\mu_{\rm eff}$ is extremely small in some simple extensions of the SM, it could be sufficiently large in some more complicated or exotic scenarios beyond the SM, such as a class of extra-dimension models. Experimentally, radiative decays of massive neutrinos can be constrained by seeing no emission of the photons from solar ν_e and reactor $\overline{\nu}_e$ fluxes. Much stronger constraints on $\mu_{\rm eff}$ can be obtained from the Supernova 1987A limit on the neutrino decay and from the astrophysical limit on distortions of the cosmic microwave background (CMB) radiation. A brief summary of these limits is

$$\frac{\mu_{\rm eff}}{\mu_{\rm B}} < \begin{cases} 0.9 \times 10^{-1} \left(\dfrac{\rm eV}{m_\nu}\right)^2 & \text{Reactor} \\ 0.5 \times 10^{-5} \left(\dfrac{\rm eV}{m_\nu}\right)^2 & \text{Sun} \\ 1.5 \times 10^{-8} \left(\dfrac{\rm eV}{m_\nu}\right)^2 & \text{SN 1987A} \\ 1.0 \times 10^{-11} \left(\dfrac{\rm eV}{m_\nu}\right)^{9/4} & \text{CMB} \end{cases}$$

where m_ν denotes the effective mass of the decaying neutrino (i.e., $m_\nu = m_i$).

3.4 Electromagnetic ν_e-e scattering

In practice, the most sensitive way of probing the electromagnetic dipole moments of a massive neutrino is to measure the cross section of elastic neutrino-electron (or antineutrino-electron) scattering, which can be expressed as a sum of the contribution from the SM (σ_0) and that from the electromagnetic dipole moments of massive neutrinos (σ_μ):

$$\frac{{\rm d}\sigma}{{\rm d}T} = \frac{{\rm d}\sigma_0}{{\rm d}T} + \frac{{\rm d}\sigma_\mu}{{\rm d}T} \; , \tag{85}$$

where $T = E_e - m_e$ denotes the kinetic energy of the recoil electron in this process. We have

$$\frac{{\rm d}\sigma_0}{{\rm d}T} = \frac{G_{\rm F}^2 m_e}{2\pi} \left[g_+^2 + g_-^2 \left(1 - \frac{T}{E_\nu} \right)^2 - g_+ g_- \frac{m_e T}{E_\nu^2} \right] \tag{86}$$

for neutrino-electron scattering, where $g_+ = 2 \sin^2 \theta_{\rm w} + 1$ for ν_e, $g_+ = 2 \sin^2 \theta_{\rm w} - 1$ for ν_μ and ν_τ, and $g_- = 2 \sin^2 \theta_{\rm w}$ for all flavors. Note that Eq. (86) is also valid for antineutrino-electron scattering if one simply exchanges the positions of g_+ and g_-. On the other hand,

$$\frac{{\rm d}\sigma_\mu}{{\rm d}T} = \frac{\alpha^2 \pi}{m_e^2} \left(\frac{1}{T} - \frac{1}{E_\nu} \right) \left(\frac{\mu_\nu}{\mu_{\rm B}} \right)^2 \tag{87}$$

with $\mu_\nu^2 \equiv |\mu_{ii}^{\rm D}|^2 + |\epsilon_{ii}^{\rm D}|^2$ (for $i = 1$, 2 or 3), which holds for both neutrinos and antineutrinos. In obtaining Eqs. (86) and (87) one has assumed the scattered neutrino to be a Dirac particle and omitted the effects of finite neutrino masses and flavor mixing (i.e., $\nu_e = \nu_1$, $\nu_\mu = \nu_2$ and $\nu_\tau = \nu_3$ have been taken). Hence there is no interference between the contributions coming from the SM and electromagnetic dipole moments — the latter leads to a helicity flip of the neutrino but the former is always helicity-conserving. While an interference term will appear if one takes account of neutrino masses and flavor mixing, its magnitude linearly depends on the neutrino masses and thus is strongly suppressed in comparison with the pure weak and electromagnetic terms. So the incoherent sum of ${\rm d}\sigma_0/{\rm d}T$ and ${\rm d}\sigma_\mu/{\rm d}T$ in Eq. (85) is actually an excellent approximation of ${\rm d}\sigma/{\rm d}T$.

It is obvious that the two terms of ${\rm d}\sigma/{\rm d}T$ depend on the kinetic energy of the recoil electron in quite different ways. In particular, ${\rm d}\sigma_\mu/{\rm d}T$ grows rapidly with decreasing values of T. Hence a measurement of smaller T can probe smaller μ_ν in this kind of experiments. The magnitude of ${\rm d}\sigma_\mu/{\rm d}T$ becomes larger than that of ${\rm d}\sigma_0/{\rm d}T$ if the condition

$$T \leq \frac{\alpha^2 \pi^2}{G_{\rm F}^2 m_e^3} \left(\frac{\mu_\nu}{\mu_{\rm B}} \right)^2 \approx 3 \times 10^{22} \left(\frac{\mu_\nu}{\mu_{\rm B}} \right)^2 {\rm keV} \tag{88}$$

is roughly satisfied, as one can easily see from Eqs. (86) and (87). No distortion of the recoil electron energy spectrum of $\nu_\alpha e^-$ or $\overline{\nu}_\alpha e^-$ scattering (for $\alpha = e, \mu, \tau$) has so far been observed in any direct laboratory experiments, and thus only the upper bounds on μ_ν can be derived. For instance, an analysis of the T-spectrum in the Super-Kamiokande experiment yields $\mu_\nu < 1.1 \times 10^{-10} \mu_{\rm B}$. More stringent bounds on μ_ν can hopefully be achieved in the future.

In view of current experimental data on neutrino oscillations, we know that neutrinos are actually massive. Hence the effects of finite neutrino masses and flavor mixing should be taken into account in calculating the cross section of elastic neutrino-electron or antineutrino-electron scattering. Here let us illustrate how the neutrino oscillation may affect the weak and electromagnetic terms of elastic $\overline{\nu}_e e^-$ scattering in a reactor experiment, where the antineutrinos are produced from the beta decay of fission products and detected by their elastic scattering with electrons in a detector. The antineutrino state created in this beta decay (via $W^- \to e^- + \overline{\nu}_e$) at the reactor is a superposition of three antineutrino mass eigenstates:

$$|\overline{\nu}_e(0)\rangle = \sum_{j=1}^{3} V_{ej} |\overline{\nu}_j\rangle \; . \tag{89}$$

Such a $\overline{\nu}_e$ beam propagates over the distance L to the detector,

$$|\overline{\nu}_e(L)\rangle = \sum_{j=1}^{3} e^{iq_j L} V_{ej} |\overline{\nu}_j\rangle \; , \tag{90}$$

in which $q_j = \sqrt{E_\nu^2 - m_j^2}$ is the momentum of ν_j with E_ν being the beam energy and m_j being the mass of ν_j. After taking account of the effect of neutrino oscillations, one obtains the differential cross section of elastic antineutrino-electron scattering as follows:

$$\frac{\mathrm{d}\sigma'}{\mathrm{d}T} = \frac{\mathrm{d}\sigma'_0}{\mathrm{d}T} + \frac{\mathrm{d}\sigma'_\mu}{\mathrm{d}T} \; , \tag{91}$$

where

$$\begin{aligned} \frac{\mathrm{d}\sigma'_0}{\mathrm{d}T} = \frac{G_{\rm F}^2 m_e}{2\pi} \Bigg\{ & g_-^2 + \left(g_- - 1\right)^2 \left(1 - \frac{T}{E_\nu}\right)^2 - g_- \left(g_- - 1\right) \frac{m_e T}{E_\nu^2} \\ & + 2 g_- \left| \sum_{j=1}^{3} e^{iq_j L} |V_{ej}|^2 \right|^2 \left[2 \left(1 - \frac{T}{E_\nu}\right)^2 - \frac{m_e T}{E_\nu^2} \right] \Bigg\} \end{aligned} \tag{92}$$

with $g_- = 2 \sin^2 \theta_{\rm w}$ for $\overline{\nu}_e$, and

$$\frac{\mathrm{d}\sigma'_\mu}{\mathrm{d}T} = \frac{\alpha^2 \pi}{m_e^2} \sum_{k=1}^{3} \left| \sum_{j=1}^{3} e^{iq_j L} V_{ej} \frac{\epsilon_{jk} + i\mu_{jk}}{\mu_{\rm B}} \right|^2 \times \left(\frac{1}{T} - \frac{1}{E_\nu} \right) \tag{93}$$

with μ_{jk} and ϵ_{jk} being the magnetic and electric transition dipole moments between ν_j and ν_k neutrinos as defined in Eq. (79). Because different neutrino mass eigenstates are in principle distinguishable in the electromagnetic $\overline{\nu}_e e^-$ scattering, their contributions to the total cross section are incoherent. Eq. (93) shows that it is in general difficult to determine or constrain the magnitudes of μ_{jk} and ϵ_{jk} (for $j, k = 1, 2, 3$) from a single measurement.

4 Lepton Flavor Mixing and CP Violation

Regardless of the dynamical origin of tiny neutrino masses [7], we may discuss lepton flavor mixing by taking account of the effective mass terms of charged leptons and Majorana neutrinos at low energies [8],

$$-\mathcal{L}'_{\rm lepton} = \overline{(e\ \mu\ \tau)_{\rm L}}\, M_l \begin{pmatrix} e \\ \mu \\ \tau \end{pmatrix}_{\rm R} + \frac{1}{2} \overline{\left(\nu_e\ \nu_\mu\ \nu_\tau\right)_{\rm L}}\, M_\nu \begin{pmatrix} \nu_e^c \\ \nu_\mu^c \\ \nu_\tau^c \end{pmatrix}_{\rm R} + {\rm h.c.} \; . \tag{94}$$

The phenomenon of lepton flavor mixing arises from a mismatch between the diagonalizations of M_l and M_ν in an arbitrary flavor basis: $V_l^\dagger M_l U_l = {\rm Diag}\{m_e, \mu_\mu, m_\tau\}$ and $V_\nu^\dagger M_\nu V_\nu^* = {\rm Diag}\{m_1, m_2, m_3\}$, where V_l, U_l and V_ν are the 3×3 unitary matrices. In the basis of mass eigenstates, it is the unitary matrix $V = V_l^\dagger V_\nu$ that will appear in the weak charged-current interactions in Eq. (12). Although the basis of $M_l = {\rm Diag}\{m_e, m_\mu, m_\tau\}$ with $V_l = \mathbf{1}$ and $V = V_\nu$ is often chosen in neutrino phenomenology, one should keep in mind that both the charged-lepton and neutrino sectors may in general contribute to lepton flavor mixing. In other words, both V_l and V_ν are not fully physical, and only their product $V = V_l^\dagger V_\nu$ is a physical description of lepton flavor mixing and CP violation at low energies.

[7] For simplicity, here we do not consider possible non-unitarity of the 3×3 neutrino mixing matrix because its effects are either absent or very small.

[8] As for Dirac neutrinos, the corresponding mass term is the same as that given in Eq. (7). In this case the neutrino mass matrix M_ν is in general not symmetric and can be diagonalized by means of the transformation $V_\nu^\dagger M_\nu U_\nu = {\rm Diag}\{m_1, m_2, m_3\}$, where both V_ν and U_ν are unitary.

4.1 Parametrizations of V

Flavor mixing among n different lepton families can be described by an $n \times n$ unitary matrix V, whose number of independent parameters relies on the nature of neutrinos. If neutrinos are Dirac particles, one may make use of $n(n-1)/2$ rotation angles and $(n-1)(n-2)/2$ phase angles to parametrize V. If neutrinos are Majorana particles, however, a full parametrization of V needs $n(n-1)/2$ rotation angles and the same number of phase angles [9]. The flavor mixing between charged leptons and Dirac neutrinos is completely analogous to that of quarks, for which a number of different parametrizations have been proposed and classified in the literature. Here we classify all possible parametrizations for the flavor mixing between charged leptons and Majorana neutrinos with $n = 3$. Regardless of the freedom of phase reassignments, we find that there are nine structurally different parametrizations for the 3×3 lepton flavor mixing matrix V.

The 3×3 lepton flavor mixing matrix V, which is often called the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix, can be expressed as a product of three unitary matrices O_1, O_2 and O_3. They correspond to simple rotations in the complex (1,2), (2,3) and (3,1) planes:

$$O_1 = \begin{pmatrix} c_1 e^{i\alpha_1} & s_1 e^{-i\beta_1} & 0 \\ -s_1 e^{i\beta_1} & c_1 e^{-i\alpha_1} & 0 \\ 0 & 0 & e^{i\gamma_1} \end{pmatrix} ,$$

$$O_2 = \begin{pmatrix} e^{i\gamma_2} & 0 & 0 \\ 0 & c_2 e^{i\alpha_2} & s_2 e^{-i\beta_2} \\ 0 & -s_2 e^{i\beta_2} & c_2 e^{-i\alpha_2} \end{pmatrix} ,$$

$$O_3 = \begin{pmatrix} c_3 e^{i\alpha_3} & 0 & s_3 e^{-i\beta_3} \\ 0 & e^{i\gamma_3} & 0 \\ -s_3 e^{i\beta_3} & 0 & c_3 e^{-i\alpha_3} \end{pmatrix} , \tag{95}$$

where $s_i \equiv \sin\theta_i$ and $c_i \equiv \cos\theta_i$ (for $i = 1, 2, 3$). Obviously $O_i O_i^\dagger = O_i^\dagger O_i = \mathbf{1}$ holds, and any two rotation matrices do not commute with each other. We find twelve different ways to arrange the product of O_1, O_2 and O_3, which can cover the whole 3×3 space and provide a full description of V. Explicitly, six of the twelve different combinations of O_i belong to the type

$$V = O_i(\theta_i, \alpha_i, \beta_i, \gamma_i) \otimes O_j(\theta_j, \alpha_j, \beta_j, \gamma_j) \otimes O_i(\theta_i', \alpha_i', \beta_i', \gamma_i') \tag{96}$$

with $i \neq j$, where the complex rotation matrix O_i occurs twice; and the other six belong to the type

$$V = O_i(\theta_i, \alpha_i, \beta_i, \gamma_i) \otimes O_j(\theta_j, \alpha_j, \beta_j, \gamma_j) \otimes O_k(\theta_k, \alpha_k, \beta_k, \gamma_k) \tag{97}$$

with $i \neq j \neq k$, in which the rotations take place in three different complex planes. The products $O_i O_j O_i$ and $O_i O_k O_i$ (for $i \neq k$) in Eq. (97) are correlated with each other, if the relevant phase parameters are switched off. Hence only nine of the twelve parametrizations, three from Eq. (96) and six from Eq. (97), are structurally different.

In each parametrization of V, there apparently exist nine phase parameters. Some of them or their combinations can be absorbed by redefining the relevant phases of charged-lepton and neutrino fields. If neutrinos are Dirac particles, V contains only a single irremovable CP-violating phase δ. If neutrinos are Majorana particles, however, there is no freedom to rearrange the relative phases of three Majorana neutrino fields. Hence V may in general contain three irremovable CP-violating phases in the Majorana case (δ and two Majorana phases). Both CP- and T-violating effects in neutrino oscillations depend only upon the Dirac-like phase δ.

[9] No matter whether neutrinos are Dirac or Majorana particles, the $n \times n$ unitary flavor mixing matrix has $(n-1)^2(n-2)^2/4$ Jarlskog invariants of CP violation defined as $\mathcal{J}_{\alpha\beta}^{ij} \equiv \mathrm{Im}\left(V_{\alpha i} V_{\beta j} V_{\alpha j}^* V_{\beta i}^*\right)$.

Different parametrizations of V are mathematically equivalent, so adopting any of them does not directly point to physical significance. But it is very likely that one particular parametrization is more useful and transparent than the others in studying the neutrino phenomenology and (or) exploring the underlying dynamics responsible for lepton mass generation and CP violation. Here we highlight two particular parametrizations of the PMNS matrix V. The first one is the so-called "standard" parametrization advocated by the Particle Data Group:

$$V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{pmatrix} \begin{pmatrix} c_{13} & 0 & s_{13} e^{-i\delta} \\ 0 & 1 & 0 \\ -s_{13} e^{i\delta} & 0 & c_{13} \end{pmatrix} \begin{pmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{pmatrix} P' \; , \tag{98}$$

where $c_{ij} \equiv \cos\theta_{ij}$ and $s_{ij} \equiv \sin\theta_{ij}$ (for $ij = 12, 13, 23$) together with the Majorana phase matrix $P' = {\rm Diag}\{e^{i\rho}, e^{i\sigma}, 1\}$. Without loss of generality, the three mixing angles $(\theta_{12}, \theta_{13}, \theta_{23})$ can all be arranged to lie in the first quadrant. Arbitrary values between 0 and 2π are allowed for three CP-violating phases (δ, ρ, σ). A remarkable merit of this parametrization is that its three mixing angles are approximately equivalent to the mixing angles of solar (θ_{12}), atmospheric (θ_{23}) and CHOOZ reactor (θ_{13}) neutrino oscillation experiments. Another useful parametrization is the Fritzsch-Xing (FX) parametrization proposed originally for quark mixing and later for lepton mixing:

$$V = \begin{pmatrix} c_l & s_l & 0 \\ -s_l & c_l & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} e^{-i\phi} & 0 & 0 \\ 0 & c & s \\ 0 & -s & c \end{pmatrix} \begin{pmatrix} c_\nu & -s_\nu & 0 \\ s_\nu & c_\nu & 0 \\ 0 & 0 & 1 \end{pmatrix} P' \; , \tag{99}$$

where $c_{l,\nu} \equiv \cos\theta_{l,\nu}$, $s_{l,\nu} \equiv \sin\theta_{l,\nu}$, $c \equiv \cos\theta$, $s \equiv \sin\theta$, and P' is a diagonal phase matrix containing two nontrivial CP-violating phases. Although the form of V in Eq. (99) is apparently different from that in Eq. (98), their corresponding flavor mixing angles $(\theta_l, \theta_\nu, \theta)$ and $(\theta_{12}, \theta_{13}, \theta_{23})$ have quite similar meanings in interpreting the experimental data on neutrino oscillations. In the limit $\theta_l = \theta_{13} = 0$, one easily arrives at $\theta_\nu = \theta_{12}$ and $\theta = \theta_{23}$. As a natural consequence of very small θ_l, three mixing angles of the FX parametrization can also be related to those of solar (θ_ν), atmospheric (θ) and CHOOZ reactor $(\theta_l \sin\theta)$ neutrino oscillation experiments in the leading-order approximation. A striking merit of this parametrization is that its six parameters have very simple renormalization-group equations when they run from a superhigh-energy scale to the electroweak scale or vice versa.

4.2 Democratic or tri-bimaximal mixing?

Current neutrino oscillation data indicate the essential feature of lepton flavor mixing: two mixing angles are quite large ($\theta_{12} \sim 34^\circ$ and $\theta_{23} \sim 45^\circ$) while the third one is very small ($\theta_{13} < 10^\circ$). Such a flavor mixing pattern is far beyond the original imagination of most people because it is rather different from the well-known quark mixing pattern ($\vartheta_{12} \approx 14.5^\circ$, $\vartheta_{23} \approx 2.6^\circ$, $\vartheta_{13} \approx 0.23^\circ$ and $\delta = 76.5^\circ$) described by the same parametrization of the Cabibbo-Kobayashi-Maskawa (CKM) matrix. To understand this difference, a number of constant lepton mixing patterns have been proposed as the starting point of model building. Possible flavor symmetries and their spontaneous or explicit breaking mechanisms hidden in those constant patterns might finally help us pin down the dynamics responsible for lepton mass generation and flavor mixing. To illustrate, let us first comment on the "democratic" neutrino mixing pattern and then pay more attention to the "tri-bimaximal" neutrino mixing pattern.

The "democratic" lepton flavor mixing pattern

$$U_0 = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{\sqrt{2}}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix} \tag{100}$$

was originally obtained by Fritzsch and Xing as the leading term of the 3×3 lepton mixing matrix from the breaking of flavor democracy or $S(3)_{\rm L} \times S(3)_{\rm R}$ symmetry of the charged-lepton mass matrix

in the basis where the Majorana neutrino mass matrix is diagonal and possesses the $S(3)$ symmetry. Its naive predictions $\theta_{12} = 45^\circ$ and $\theta_{23} \approx 54.7^\circ$ are no more favored today, but they may receive proper corrections from the symmetry-breaking perturbations so as to fit current neutrino oscillation data.

Today's most popular constant pattern of neutrino mixing is the "tri-bimaximal" mixing matrix:

$$V_0 = \begin{pmatrix} \frac{\sqrt{2}}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 \\ \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \end{pmatrix} \tag{101}$$

which looks like a twisted form of the democratic mixing pattern with the same entries. Its strange name comes from the fact that this flavor mixing pattern is actually a product of the "tri-maximal" mixing matrix and a "bi-maximal" mixing matrix:

$$V_0' = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{\omega}{\sqrt{3}} & \frac{\omega^2}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{\omega^2}{\sqrt{3}} & \frac{\omega}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix} = P V_0 P' \; , \tag{102}$$

where $\omega = e^{i2\pi/3}$ denotes the complex cube-root of unity (i.e., $\omega^3 = 1$), and $P = {\rm Diag}\{1, \omega, \omega^2\}$ and $P' = {\rm Diag}\{1, 1, i\}$ are two diagonal phase matrices. V_0 or V_0' predicts $\theta_{12} = \arctan(1/\sqrt{2}) \approx 35.3^\circ$, $\theta_{13} = 0^\circ$ and $\theta_{23} = 45^\circ$, consistent quite well with current neutrino oscillation data. Because the entries of U_0 or V_0 are all formed from small integers (0, 1, 2 and 3) and their square roots, it is often suggestive of certain discrete flavor symmetries in the language of group theories. That is why the democratic or tri-bimaximal neutrino mixing pattern can serve as a good starting point of model building based on a variety of flavor symmetries, such as Z_2, Z_3, S_3, S_4, A_4, D_4, D_5, Q_4, Q_6, $\Delta(27)$ and $\Sigma(81)$. In particular, a lot of interest has been paid to the derivation of V_0 with the help of the non-Abelian discrete A_4 symmetry.

Note that the democratic mixing matrix U_0 and the tri-bimaximal mixing matrix V_0 are related with each other via the following transformation:

$$V_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_0 & -\sin\theta_0 \\ 0 & \sin\theta_0 & \cos\theta_0 \end{pmatrix} U_0 \begin{pmatrix} \cos\theta_0 & -\sin\theta_0 & 0 \\ \sin\theta_0 & \cos\theta_0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \; , \tag{103}$$

where $\theta_0 = \arctan(\sqrt{2} - 1)^2 \approx 9.7^\circ$. This angle is actually a measure of the difference between the mixing angles of U_0 and V_0 (namely, $45^\circ - 35.3^\circ = 54.7^\circ - 45^\circ = 9.7^\circ$). In this sense, we argue that it is worthwhile to explore possible flavor symmetries behind both V_0 and U_0 so as to build realistic models for neutrino mass generation and lepton flavor mixing.

Let us remark that a specific constant mixing pattern should be regarded as the leading-order approximation of the "true" lepton flavor mixing matrix, whose mixing angles should in general depend on both the ratios of charged-lepton masses and those of neutrino masses. We may at least make the following naive speculation about how to phenomenologically understand the observed pattern of lepton flavor mixing:

- Large values of θ_{12} and θ_{23} could arise from a weak hierarchy or a near degeneracy of the neutrino mass spectrum, because the strong hierarchy of charged-lepton masses implies that m_e/m_μ and m_μ/m_τ at the electroweak scale are unlikely to contribute to θ_{12} and θ_{23} in a dominant way.
- Special values of θ_{12} and θ_{23} might stem from an underlying flavor symmetry of the charged-lepton mass matrix or the neutrino mass matrix. Then the contributions of lepton mass ratios to flavor mixing angles, due to flavor symmetry breaking, are expected to serve as perturbative corrections to U_0 or V_0, or another constant mixing pattern.

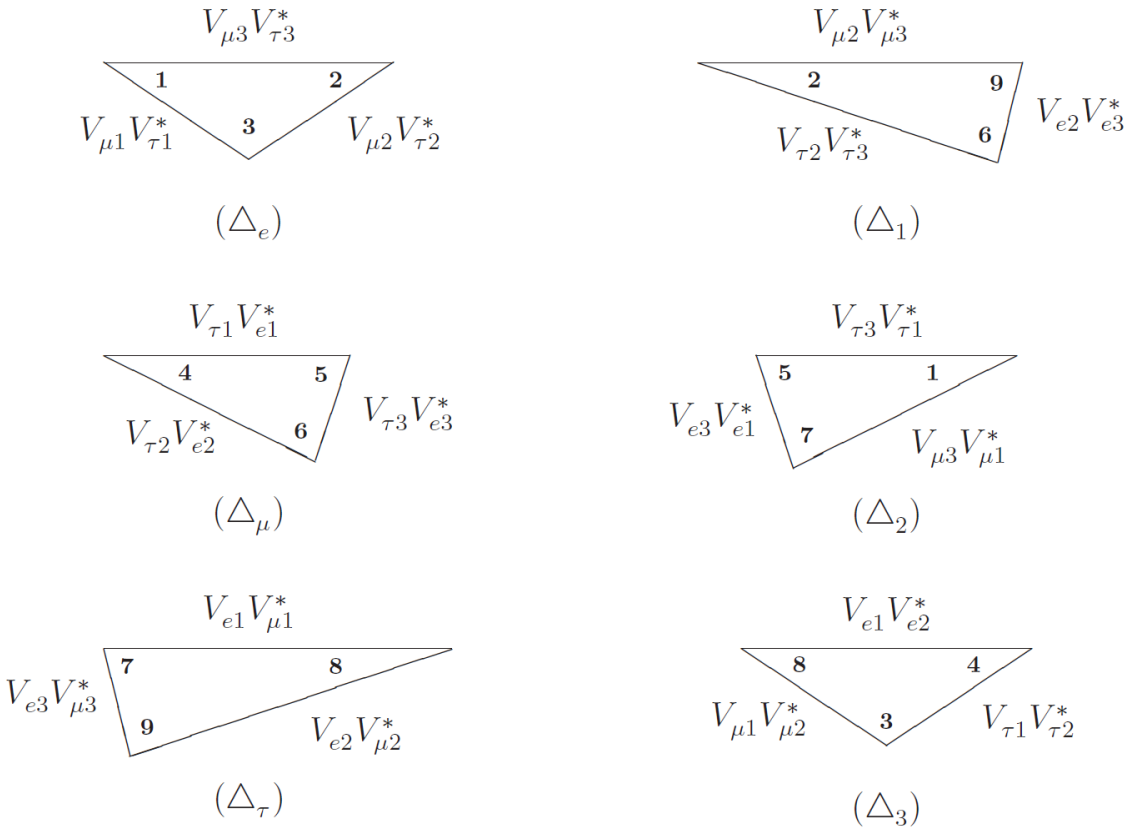


Fig. 2: Unitarity triangles of the 3×3 PMNS matrix in the complex plane. Each triangle is named by the index that does not manifest in its three sides.

– Vanishing or small θ_{13} could be a natural consequence of the explicit textures of lepton mass matrices. It might also be related to the flavor symmetry which gives rise to sizable θ_{12} and θ_{23} (e.g., in U_0 or V_0).
– Small corrections to a constant flavor mixing pattern may also result from the renormalization-group running effects of leptons and quarks, e.g., from a superhigh-energy scale to low energies or vice versa.

There are too many possibilities of linking the observed pattern of lepton flavor mixing to a certain flavor symmetry, and none of them is unique from the theoretical point of view. In this sense, flavor symmetries should not be regarded as a perfect guiding principle of model building.

4.3 Leptonic unitarity triangles

In the basis where the flavor eigenstates of charged leptons are identified with their mass eigenstates, the PMNS matrix V relates the neutrino mass eigenstates (ν_1, ν_2, ν_3) to the neutrino flavor eigenstates $(\nu_e, \nu_\mu, \nu_\tau)$:

$$\begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix} = \begin{pmatrix} V_{e1} & V_{e2} & V_{e3} \\ V_{\mu1} & V_{\mu2} & V_{\mu3} \\ V_{\tau1} & V_{\tau2} & V_{\tau3} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} \; . \tag{104}$$

The unitarity of V represents two sets of normalization and orthogonality conditions:

$$\sum_i \left(V_{\alpha i} V^*_{\beta i}\right) = \delta_{\alpha\beta} \; , \quad \sum_\alpha \left(V_{\alpha i} V^*_{\alpha j}\right) = \delta_{ij} \; , \tag{105}$$

where Greek and Latin subscripts run over (e, μ, τ) and $(1, 2, 3)$, respectively. In the complex plane the six orthogonality relations in Eq. (105) define six triangles $(\triangle_e, \triangle_\mu, \triangle_\tau)$ and $(\triangle_1, \triangle_2, \triangle_3)$ shown in Fig. 2, the so-called unitarity triangles. These six triangles have eighteen different sides and nine different inner (or outer) angles. But the unitarity of V requires that all six triangles have the same area amounting to $\mathcal{J}/2$, where $\mathcal{J}$ is the Jarlskog invariant of CP violation defined through

$$\mathrm{Im}\left(V_{\alpha i} V_{\beta j} V^*_{\alpha j} V^*_{\beta i}\right) = \mathcal{J} \sum_\gamma \epsilon_{\alpha\beta\gamma} \sum_k \epsilon_{ijk} \; . \tag{106}$$

One has $\mathcal{J} = c_{12} s_{12} c^2_{13} s_{13} c_{23} s^2_{23} \sin\delta$ in the standard parametrization of V as well as $\mathcal{J} = c_l s_l c_\nu s_\nu c s^2 \sin\phi$ in the FX parametrization of V. No matter whether neutrinos are Dirac or Majorana particles, the strength of CP or T violation in neutrino oscillations depends only upon $\mathcal{J}$.

To show why the areas of six unitarity triangles are identical with one another, let us take triangles $\triangle_\tau$ and $\triangle_3$ for example. They correspond to the orthogonality relations

$$\begin{aligned} V_{e1} V^*_{\mu 1} + V_{e2} V^*_{\mu 2} + V_{e3} V^*_{\mu 3} &= 0 \; , \\ V_{e1} V^*_{e 2} + V_{\mu 1} V^*_{\mu 2} + V_{\tau 1} V^*_{\tau 2} &= 0 \; . \end{aligned} \tag{107}$$

Multiplying these two equations by $V_{\mu 2} V^*_{e 2}$ and $V_{\mu 2} V^*_{\mu 1}$ respectively, we arrive at two rescaled triangles which share the side

$$V_{e1} V_{\mu 2} V^*_{e 2} V^*_{\mu 1} = -|V_{e2} V_{\mu 2}|^2 - V_{e3} V_{\mu 2} V^*_{e 2} V^*_{\mu 3} = -|V_{\mu 1} V_{\mu 2}|^2 - V_{\mu 2} V_{\tau 1} V^*_{\mu 1} V^*_{\tau 2} \; . \tag{108}$$

This result is consistent with the definition of $\mathcal{J}$ in Eq. (106); i.e., $\mathrm{Im}(V_{e1} V_{\mu 2} V^*_{e 2} V^*_{\mu 1}) = \mathcal{J}$ and $\mathrm{Im}(V_{e3} V_{\mu 2} V^*_{e 2} V^*_{\mu 3}) = \mathrm{Im}(V_{\mu 2} V_{\tau 1} V^*_{\mu 1} V^*_{\tau 2}) = -\mathcal{J}$. The latter simultaneously implies that the areas of $\triangle_\tau$ and $\triangle_3$ are equal to $\mathcal{J}/2$. One may analogously prove that all the six unitarity triangles have the same area $\mathcal{J}/2$. If CP or T were an exact symmetry, $\mathcal{J} = 0$ would hold and those unitarity triangles would collapse into lines in the complex plane. Note that the shape and area of each unitarity triangle are irrelevant to the nature of neutrinos; i.e., they are the same for Dirac and Majorana neutrinos.

Because of $V^*_{e1} V_{\mu 1} + V^*_{e 2} V_{\mu 2} = -V^*_{e 3} V_{\mu 3}$ or equivalently $|V_{e1} V^*_{\mu 1} + V_{e2} V^*_{\mu 2}|^2 = |V_{e3} V^*_{\mu 3}|^2$, it is easy to obtain

$$2\mathrm{Re}\left(V_{e1} V_{\mu 2} V^*_{e 2} V^*_{\mu 1}\right) = |V_{e3}|^2 |V_{\mu 3}|^2 - |V_{e1}|^2 |V_{\mu 1}|^2 - |V_{e2}|^2 |V_{\mu 2}|^2 \; . \tag{109}$$

Combining $V_{e1} V_{\mu 2} V^*_{e 2} V^*_{\mu 1} = \mathrm{Re}(V_{e1} V_{\mu 2} V^*_{e 2} V^*_{\mu 1}) + i\mathcal{J}$ with Eq. (109) leads us to the result

$$\begin{aligned} \mathcal{J}^2 &= |V_{e1}|^2 |V_{\mu 2}|^2 |V_{e2}|^2 |V_{\mu 1}|^2 - \frac{1}{4} \left(|V_{e3}|^2 |V_{\mu 3}|^2 - |V_{e1}|^2 |V_{\mu 1}|^2 - |V_{e2}|^2 |V_{\mu 2}|^2\right)^2 \\ &= |V_{e1}|^2 |V_{\mu 2}|^2 |V_{e2}|^2 |V_{\mu 1}|^2 - \frac{1}{4} \left(1 + |V_{e1}|^2 |V_{\mu 2}|^2 + |V_{e2}|^2 |V_{\mu 1}|^2 \right. \\ &\qquad \left. - |V_{e1}|^2 - |V_{\mu 2}|^2 - |V_{e2}|^2 - |V_{\mu 1}|^2\right)^2 \; . \end{aligned} \tag{110}$$

As a straightforward generalization of Eq. (110), $\mathcal{J}^2$ can be expressed in terms of the moduli of any four independent matrix elements of V:

$$\begin{aligned} \mathcal{J}^2 = |V_{\alpha i}|^2 |V_{\beta j}|^2 |V_{\alpha j}|^2 |V_{\beta i}|^2 - \frac{1}{4} \left(1 + |V_{\alpha i}|^2 |V_{\beta j}|^2 + |V_{\alpha j}|^2 |V_{\beta i}|^2 \right. \\ \left. - |V_{\alpha i}|^2 - |V_{\beta j}|^2 - |V_{\alpha j}|^2 - |V_{\beta i}|^2\right)^2 \; , \end{aligned} \tag{111}$$

in which $\alpha \neq \beta$ running over (e, μ, τ) and $i \neq j$ running over $(1, 2, 3)$. The implication of this result is very obvious: the information about leptonic CP violation can in principle be extracted from the measured moduli of the neutrino mixing matrix elements.

As a consequence of the unitarity of V, two interesting relations can be derived from the normalization conditions in Eq. (105):

$$|V_{e2}|^2 - |V_{\mu 1}|^2 = |V_{\mu 3}|^2 - |V_{\tau 2}|^2 = |V_{\tau 1}|^2 - |V_{e3}|^2 \equiv \Delta_{\rm L} \; ,$$
$$|V_{e2}|^2 - |V_{\mu 3}|^2 = |V_{\mu 1}|^2 - |V_{\tau 2}|^2 = |V_{\tau 3}|^2 - |V_{e1}|^2 \equiv \Delta_{\rm R} \; . \tag{112}$$

The off-diagonal asymmetries $\Delta_{\rm L}$ and $\Delta_{\rm R}$ characterize the geometrical structure of V about its V_{e1}-$V_{\mu 2}$-$V_{\tau 3}$ and V_{e3}-$V_{\mu 2}$-$V_{\tau 1}$ axes, respectively. For instance, $\Delta_{\rm L} = 1/6$ and $\Delta_{\rm R} = -1/6$ hold for the tri-bimaximal neutrino mixing pattern V_0. If $\Delta_{\rm L} = 0$ (or $\Delta_{\rm R} = 0$) held, V would be symmetric about the V_{e1}-$V_{\mu 2}$-$V_{\tau 3}$ (or V_{e3}-$V_{\mu 2}$-$V_{\tau 1}$) axis. Geometrically this would correspond to the congruence between two unitarity triangles; i.e.,

$$\Delta_{\rm L} = 0 : \; \triangle_e \cong \triangle_1 \; , \triangle_\mu \cong \triangle_2 \; , \triangle_\tau \cong \triangle_3 \; ;$$
$$\Delta_{\rm R} = 0 : \; \triangle_e \cong \triangle_3 \; , \triangle_\mu \cong \triangle_2 \; , \triangle_\tau \cong \triangle_1 \; . \tag{113}$$

Indeed the counterpart of $\Delta_{\rm L}$ in the quark sector is only of $\mathcal{O}(10^{-5})$; i.e., the CKM matrix is almost symmetric about its V_{ud}-V_{cs}-V_{tb} axis. An exactly symmetric flavor mixing matrix might hint at an underlying flavor symmetry, from which some deeper understanding of the fermion mass texture could be achieved.

4.4 Flavor problems in particle physics

In the subatomic world the fundamental building blocks of matter have twelve flavors: six quarks and six leptons (and their antiparticles). Table 2 is a brief list of some important discoveries in flavor physics, which can partly give people a ball-park feeling of a century of developments in particle physics. The SM of electromagnetic and weak interactions contain thirteen free parameters in its lepton and quark sectors: three charged-lepton masses, six quark masses, three quark flavor mixing angles and one CP-violating phase. If three known neutrinos are massive Majorana particles, one has to introduce nine free parameters to describe their flavor properties: three neutrino masses, three lepton flavor mixing angles and three CP-violating phases. Thus an effective theory of electroweak interactions at low energies totally consists of twenty-two flavor parameters which can only be determined from experiments. Why is the number of degrees of freedom so big in the flavor sector? What is the fundamental physics behind these parameters? Such puzzles constitute the flavor problems in particle physics.

Current experimental data on neutrino oscillations can only tell us $m_1 < m_2$. It remains unknown whether m_3 is larger than m_2 (normal hierarchy) or smaller than m_1 (inverted hierarchy). The possibility $m_1 \approx m_2 \approx m_3$ (near degeneracy) cannot be excluded at present. In contrast, three families of charged fermions have very strong mass hierarchies:

$$\frac{m_e}{m_\mu} \sim \frac{m_u}{m_c} \sim \frac{m_c}{m_t} \sim \lambda^4 \; ,$$
$$\frac{m_\mu}{m_\tau} \sim \frac{m_d}{m_s} \sim \frac{m_s}{m_b} \sim \lambda^2 \; , \tag{114}$$

where $\lambda \equiv \sin\theta_{\rm C} \approx 0.22$ with $\theta_{\rm C}$ being the Cabibbo angle of quark flavor mixing. In the standard parametrization of the CKM matrix, three quark mixing angles exhibit an impressive hierarchy:

$$\vartheta_{12} \sim \lambda \; , \quad \vartheta_{23} \sim \lambda^2 \; , \quad \vartheta_{13} \sim \lambda^4 \; . \tag{115}$$

These two kinds of hierarchies might intrinsically be related to each other, because the flavor mixing angles actually measure a mismatch between the mass and flavor eigenstates of up- and down-type quarks. For example, the relations $\vartheta_{12} \approx \sqrt{m_d/m_s}$, $\vartheta_{23} \approx \sqrt{m_d/m_b}$ and $\vartheta_{13} \approx \sqrt{m_u/m_t}$ are compatible with Eqs. (114) and (115). They can be derived from a specific pattern of up- and down-type quark

Table 2: Some important discoveries in the developments of flavor physics.

	Discoveries of lepton flavors, quark flavors and CP violation
1897	electron (Thomson, 1897)
1919	proton (up and down quarks) (Rutherford, 1919)
1932	neutron (up and down quarks) (Chadwick, 1932)
1933	positron (Anderson, 1933)
1936	muon (Neddermeyer and Anderson, 1937)
1947	Kaon (strange quark) (Rochester and Butler, 1947)
1956	electron antineutrino (Cowan *et al.*, 1956)
1962	muon neutrino (Danby *et al.*, 1962)
1964	CP violation in s-quark decays (Christenson *et al.*, 1964)
1974	charm quark (Aubert *et al.*, 1974; Abrams *et al.*, 1974)
1975	tau (Perl *et al.*, 1975)
1977	bottom quark (Herb *et al.*, 1977)
1995	top quark (Abe *et al.*, 1995; Abachi *et al.*, 1995)
2000	tau neutrino (Kodama *et al.*, 2000)
2001	CP violation in b-quark decays (Aubert *et al.*, 2001; Abe *et al.*, 2001)

mass matrices with five texture zeros. On the other hand, it seems quite difficult to find a simple way of linking two large lepton flavor mixing angles $\theta_{12} \sim \pi/6$ and $\theta_{23} \sim \pi/4$ to small m_e/m_μ and m_μ/m_τ. One might ascribe the largeness of θ_{12} and θ_{23} to a very weak hierarchy of three neutrino masses and the smallness of θ_{13} to the strong mass hierarchy in the charged-lepton sector. There are of course many possibilities of model building to understand the observed lepton flavor mixing pattern, but none of them has experimentally and theoretically been justified.

Among a number of concrete flavor puzzles that are currently facing us, the following three are particularly intriguing.

– The pole masses of three charged leptons satisfy the equality

$$\frac{m_e + m_\mu + m_\tau}{\left(\sqrt{m_e} + \sqrt{m_\mu} + \sqrt{m_\tau}\right)^2} = \frac{2}{3} \tag{116}$$

to an amazingly good degree of accuracy — its error bar is only of $\mathcal{O}(10^{-5})$.

– There are two quark-lepton "complementarity" relations in flavor mixing:

$$\theta_{12} + \vartheta_{12} \approx \theta_{23} + \vartheta_{23} \approx \frac{\pi}{4} \; , \tag{117}$$

which are compatible with the present experimental data.

– Two unitarity triangles of the CKM matrix, defined by the orthogonality conditions $V_{ud}V^*_{ub} + V_{cd}V^*_{cb} + V_{td}V^*_{tb} = 0$ and $V_{tb}V^*_{ub} + V_{ts}V^*_{us} + V_{td}V^*_{ud} = 0$, are almost the right triangles. Namely, the common inner angle of these two triangles satisfies

$$\alpha \equiv \arg\left(-\frac{V_{ud}V^*_{ub}}{V_{td}V^*_{tb}}\right) \approx \frac{\pi}{2} \; , \tag{118}$$

indicated by current experimental data on quark mixing and CP violation.

Such special numerical relations might just be accidental. One or two of them might also be possible to result from a certain (underlying) flavor symmetry.

5 Running of Neutrino Mass Parameters

5.1 One-loop RGEs

The spirit of seesaw mechanisms is to attribute the small masses of three known neutrinos to the existence of some heavy degrees of freedom, such as the $SU(2)_{\rm L}$ gauge-singlet fermions, the $SU(2)_{\rm L}$ gauge-triplet scalars or the $SU(2)_{\rm L}$ gauge-triplet fermions. All of them point to the unique dimension-5 Weinberg operator in an effective theory after the corresponding heavy particles are integrated out:

$$\frac{\mathcal{L}_{\rm d=5}}{\Lambda} = \frac{1}{2}\kappa_{\alpha\beta}\overline{\ell_{\alpha{\rm L}}}\tilde{H}\tilde{H}^T\ell^c_{\beta{\rm L}} + {\rm h.c.}\;, \tag{119}$$

where Λ is the cutoff scale, $\ell_{\rm L}$ denotes the left-handed lepton doublet, $\tilde{H} \equiv i\sigma_2 H^*$ with H being the SM Higgs doublet, and κ stands for the effective neutrino coupling matrix. After spontaneous gauge symmetry breaking, $\tilde{H}$ gains its vacuum expectation value $\langle\tilde{H}\rangle = v/\sqrt{2}$ with $v \approx 246$ GeV. We are then left with the effective Majorana mass matrix $M_\nu = \kappa v^2/2$ for three light neutrinos from Eq. (119). If the dimension-5 Weinberg operator is obtained in the framework of the minimal supersymmetric standard model (MSSM), one will be left with $M_\nu = \kappa(v\sin\beta)^2/2$, where $\tan\beta$ denotes the ratio of the vacuum expectation values of two MSSM Higgs doublets.

Eq. (119) or its supersymmetric counterpart can provide a simple but generic way of generating tiny neutrino masses. There are a number of interesting possibilities of building renormalizable gauge models to realize the effective Weinberg mass operator, either radiatively or at the tree level. The latter case is just associated with the well-known seesaw mechanisms to be discussed in section 6. Here we assume that $\mathcal{L}_{\rm d=5}/\Lambda$ arises from an underlying seesaw model, whose lightest heavy particle has a mass of $\mathcal{O}(\Lambda)$. In other words, Λ characterizes the seesaw scale. Above Λ there may exist one or more energy thresholds corresponding to the masses of heavier seesaw particles. Below Λ the energy dependence of the effective neutrino coupling matrix κ is described by its renormalization-group equation (RGE). The evolution of κ from Λ down to the electroweak scale is formally independent of any details of the relevant seesaw model from which κ is derived.

At the one-loop level κ obeys the RGE

$$16\pi^2\frac{{\rm d}\kappa}{{\rm d}t} = \alpha_\kappa\kappa + C_\kappa\left[(Y_lY_l^\dagger)\kappa + \kappa(Y_lY_l^\dagger)^T\right] \tag{120}$$

where $t \equiv \ln(\mu/\Lambda)$ with μ being an arbitrary renormalization scale between the electroweak scale and the seesaw scale, and Y_l is the charged-lepton Yukawa coupling matrix. The RGE of Y_l and those of $Y_{\rm u}$ (up-type quarks) and $Y_{\rm d}$ (down-type quarks) are given by

$$\begin{aligned}
16\pi^2\frac{{\rm d}Y_l}{{\rm d}t} &= \left[\alpha_l + C_l^l(Y_lY_l^\dagger)\right]Y_l\;,\\
16\pi^2\frac{{\rm d}Y_{\rm u}}{{\rm d}t} &= \left[\alpha_{\rm u} + C_{\rm u}^{\rm u}(Y_{\rm u}Y_{\rm u}^\dagger) + C_{\rm u}^{\rm d}(Y_{\rm d}Y_{\rm d}^\dagger)\right]Y_{\rm u}\;,\\
16\pi^2\frac{{\rm d}Y_{\rm d}}{{\rm d}t} &= \left[\alpha_{\rm d} + C_{\rm d}^{\rm u}(Y_{\rm u}Y_{\rm u}^\dagger) + C_{\rm d}^{\rm d}(Y_{\rm d}Y_{\rm d}^\dagger)\right]Y_{\rm d}\;.
\end{aligned} \tag{121}$$

In the framework of the SM we have

$$\begin{aligned}
C_\kappa &= C_{\rm u}^{\rm d} = C_{\rm d}^{\rm u} = -\frac{3}{2}\;,\\
C_l^l &= C_{\rm u}^{\rm u} = C_{\rm d}^{\rm d} = +\frac{3}{2}\;,
\end{aligned} \tag{122}$$

and

$$\alpha_\kappa = -3g_2^2 + \lambda + 2{\rm Tr}\left[3(Y_{\rm u}Y_{\rm u}^\dagger) + 3(Y_{\rm d}Y_{\rm d}^\dagger) + (Y_lY_l^\dagger)\right]\;,$$

$$\alpha_l = -\frac{9}{4}g_1^2 - \frac{9}{4}g_2^2 + {\rm Tr}\left[3(Y_{\rm u}Y_{\rm u}^\dagger) + 3(Y_{\rm d}Y_{\rm d}^\dagger) + (Y_lY_l^\dagger)\right] \; ,$$
$$\alpha_{\rm u} = -\frac{17}{20}g_1^2 - \frac{9}{4}g_2^2 - 8g_3^2 + {\rm Tr}\left[3(Y_{\rm u}Y_{\rm u}^\dagger) + 3(Y_{\rm d}Y_{\rm d}^\dagger) + (Y_lY_l^\dagger)\right] \; ,$$
$$\alpha_{\rm d} = -\frac{1}{4}g_1^2 - \frac{9}{4}g_2^2 - 8g_3^2 + {\rm Tr}\left[3(Y_{\rm u}Y_{\rm u}^\dagger) + 3(Y_{\rm d}Y_{\rm d}^\dagger) + (Y_lY_l^\dagger)\right] \; ; \tag{123}$$

and in the framework of the MSSM we have

$$C_\kappa = C_{\rm u}^{\rm d} = C_{\rm d}^{\rm u} = +1 \; ,$$
$$C_l^l = C_{\rm u}^{\rm u} = C_{\rm d}^{\rm d} = +3 \; , \tag{124}$$

and

$$\alpha_\kappa = -\frac{6}{5}g_1^2 - 6g_2^2 + 6{\rm Tr}(Y_{\rm u}Y_{\rm u}^\dagger) \; ,$$
$$\alpha_l = -\frac{9}{5}g_1^2 - 3g_2^2 + {\rm Tr}\left[3(Y_{\rm d}Y_{\rm d}^\dagger) + (Y_lY_l^\dagger)\right] \; ,$$
$$\alpha_{\rm u} = -\frac{13}{15}g_1^2 - 3g_2^2 - \frac{16}{3}g_3^2 + 3{\rm Tr}(Y_{\rm u}Y_{\rm u}^\dagger) \; ,$$
$$\alpha_{\rm d} = -\frac{7}{15}g_1^2 - 3g_2^2 - \frac{16}{3}g_3^2 + {\rm Tr}\left[3(Y_{\rm d}Y_{\rm d}^\dagger) + (Y_lY_l^\dagger)\right] \; . \tag{125}$$

Here g_1, g_2 and g_3 are the gauge couplings and satisfy their RGEs

$$16\pi^2 \frac{{\rm d}g_i}{{\rm d}t} = b_i g_i^3 \; , \tag{126}$$

where $(b_1, b_2, b_3) = (41/10, -19/6, -7)$ in the SM or $(33/5, 1, -3)$ in the MSSM. In addition, λ is the Higgs self-coupling parameter of the SM and obeys the RGE

$$16\pi^2 \frac{{\rm d}\lambda}{{\rm d}t} = 6\lambda^2 - 3\lambda\left(\frac{3}{5}g_1^2 + 3g_2^2\right) + \frac{3}{2}\left(\frac{3}{5}g_1^2 + g_2^2\right)^2 + 3g_2^4$$
$$+4\lambda{\rm Tr}\left[3(Y_{\rm u}Y_{\rm u}^\dagger) + 3(Y_{\rm d}Y_{\rm d}^\dagger) + (Y_lY_l^\dagger)\right]$$
$$-8{\rm Tr}\left[3(Y_{\rm u}Y_{\rm u}^\dagger)^2 + 3(Y_{\rm d}Y_{\rm d}^\dagger)^2 + (Y_lY_l^\dagger)^2\right] \; . \tag{127}$$

The relation between λ and the Higgs mass M_h is given by $\lambda = M_h^2/(2v^2)$, where $v \approx 246$ GeV is the vacuum expectation value of the Higgs field.

The above RGEs allow us to evaluate the running behavior of κ together with those of Y_l, $Y_{\rm u}$ and $Y_{\rm d}$, from the seesaw scale to the electroweak scale or vice versa. We shall examine the evolution of neutrino masses, lepton flavor mixing angles and CP-violating phases in the following.

5.2 Running neutrino mass parameters

Without loss of any generality, we choose the flavor basis where Y_l is diagonal: $Y_l = D_l \equiv {\rm Diag}\{y_e, y_\mu, y_\tau\}$ with y_α being the eigenvalues of Y_l. In this case the effective Majorana neutrino coupling matrix κ can be diagonalized by the PMNS matrix V; i.e., $V^\dagger \kappa V^* = \widehat{\kappa} \equiv {\rm Diag}\{\kappa_1, \kappa_2, \kappa_3\}$ with κ_i being the eigenvalues of κ. Then

$$\frac{{\rm d}\kappa}{{\rm d}t} = \dot{V}\widehat{\kappa}V^T + V\dot{\widehat{\kappa}}V^T + V\widehat{\kappa}\dot{V}^T = \frac{1}{16\pi^2}\left[\alpha_\kappa V\widehat{\kappa}V^T + C_\kappa\left(D_l^2 V\widehat{\kappa}V^T + V\widehat{\kappa}V^T D_l^2\right)\right] \; , \tag{128}$$

with the help of Eq. (120). After a definition of the Hermitian matrix $S \equiv V^\dagger D_l^2 V$ and the anti-Hermitian matrix $T \equiv V^\dagger \dot{V}$, Eq. (128) leads to

$$\dot{\widehat{\kappa}} = \frac{1}{16\pi^2}\left[\alpha_\kappa\widehat{\kappa} + C_\kappa(S\widehat{\kappa} + \widehat{\kappa}S^*)\right] - T\widehat{\kappa} + \widehat{\kappa}T^* . \tag{129}$$

Because $\widehat{\kappa}$ is by definition diagonal and real, the left- and right-hand sides of Eq. (129) must be diagonal and real. We can therefore arrive at

$$\dot{\kappa}_i = \frac{1}{16\pi^2} \left(\alpha_\kappa + 2C_\kappa {\rm Re} S_{ii} \right) \kappa_i \; , \tag{130}$$

together with ${\rm Im}T_{ii} = {\rm Re}T_{ii} = {\rm Im}S_{ii} = 0$ (for $i = 1, 2, 3$). As the off-diagonal parts of Eq. (129) are vanishing, we have

$$T_{ij}\kappa_j - \kappa_i T^*_{ij} = \frac{C_\kappa}{16\pi^2} \left(S_{ij}\kappa_j + \kappa_i S^*_{ij} \right) \tag{131}$$

with $i \neq j$. Therefore,

$$\begin{aligned} {\rm Re}T_{ij} &= -\frac{C_\kappa}{16\pi^2} \frac{\kappa_i + \kappa_j}{\kappa_i - \kappa_j} {\rm Re}S_{ij} \; , \\ {\rm Im}T_{ij} &= -\frac{C_\kappa}{16\pi^2} \frac{\kappa_i - \kappa_j}{\kappa_i + \kappa_j} {\rm Im}S_{ij} \; . \end{aligned} \tag{132}$$

Due to $\dot{V} = VT$, Eq. (132) actually governs the evolution of V with energies.

We proceed to define $V \equiv PUP'$, in which $P \equiv {\rm Diag}\{e^{i\phi_e}, e^{i\phi_\mu}, e^{i\phi_\tau}\}$, $P' \equiv {\rm Diag}\{e^{i\rho}, e^{i\sigma}, 1\}$, and U is the CKM-like matrix containing three neutrino mixing angles and one CP-violating phase. Although P does not have any physical meaning, its phases have their own RGEs. In contrast, P' serves for the Majorana phase matrix. We find

$$T' \equiv P'TP'^\dagger = P'V^\dagger \dot{V} P'^\dagger = \dot{P}'P'^\dagger + U^\dagger \dot{U} + U^\dagger P^\dagger \dot{P} U \; , \tag{133}$$

from which we can obtain six independent constraint equations:

$$\begin{aligned} T'_{11} &= i\dot{\rho} + \sum_\alpha \left[U^*_{\alpha 1} \dot{U}_{\alpha 1} + i U_{\alpha 1} \dot{\phi}_\alpha \right] \; , \\ T'_{22} &= i\dot{\sigma} + \sum_\alpha \left[U^*_{\alpha 2} \dot{U}_{\alpha 2} + i U_{\alpha 2} \dot{\phi}_\alpha \right] \; , \\ T'_{33} &= \sum_\alpha \left[U^*_{\alpha 3} \dot{U}_{\alpha 3} + i U_{\alpha 3} \dot{\phi}_\alpha \right] \; ; \\ T'_{12} &= \sum_\alpha \left[U^*_{\alpha 1} \dot{U}_{\alpha 2} + i U_{\alpha 2} \dot{\phi}_\alpha \right] \; , \\ T'_{13} &= \sum_\alpha \left[U^*_{\alpha 1} \dot{U}_{\alpha 3} + i U_{\alpha 3} \dot{\phi}_\alpha \right] \; , \\ T'_{23} &= \sum_\alpha \left[U^*_{\alpha 2} \dot{U}_{\alpha 3} + i U_{\alpha 3} \dot{\phi}_\alpha \right] \; , \end{aligned} \tag{134}$$

where α runs over e, μ and τ. Note that $T_{ii} = 0$ holds and T_{ij} is given by Eq. (132). In view of $y_e \ll y_\mu \ll y_\tau$, we take $D^2_l \approx {\rm Diag}\{0, 0, y^2_\tau\}$ as an excellent approximation. Then S_{ij}, T_{ij} and T'_{ij} can all be expressed in terms of y^2_τ and the parameters of U and P'. After a straightforward calculation, we obtain the explicit expressions of Eqs. (130) and (134) as follows:

$$\dot{\kappa}_i = \frac{\kappa_i}{16\pi^2} \left(\alpha_\kappa + 2C_\kappa y^2_\tau |U_{\tau i}|^2 \right) \; , \tag{135}$$

and

$$\sum_\alpha \left[U^*_{\alpha 1} \left(i\dot{U}_{\alpha 1} - U_{\alpha 1} \dot{\phi}_\alpha \right) \right] = \dot{\rho} \; ,$$

$$\sum_\alpha \left[U^*_{\alpha 2}\left(i\dot{U}_{\alpha 2} - U_{\alpha 2}\dot{\phi}_\alpha\right)\right] = \dot{\sigma} \; ,$$
$$\sum_\alpha \left[U^*_{\alpha 3}\left(i\dot{U}_{\alpha 3} - U_{\alpha 3}\dot{\phi}_\alpha\right)\right] = 0 \; ,$$
$$\sum_\alpha \left[U^*_{\alpha 1}\left(\dot{U}_{\alpha 2} + iU_{\alpha 2}\dot{\phi}_\alpha\right)\right] = -\frac{C_\kappa y^2_\tau}{16\pi^2} e^{i(\rho-\sigma)} \left[\zeta^{-1}_{12}{\rm Re}\left(U^*_{\tau 1}U_{\tau 2}e^{i(\sigma-\rho)}\right) + i\zeta_{12}{\rm Im}\left(U^*_{\tau 1}U_{\tau 2}e^{i(\sigma-\rho)}\right)\right]$$
$$\sum_\alpha \left[U^*_{\alpha 1}\left(\dot{U}_{\alpha 3} + iU_{\alpha 3}\dot{\phi}_\alpha\right)\right] = -\frac{C_\kappa y^2_\tau}{16\pi^2} e^{i\rho} \left[\zeta^{-1}_{13}{\rm Re}\left(U^*_{\tau 1}U_{\tau 3}e^{-i\rho}\right) + i\zeta_{13}{\rm Im}\left(U^*_{\tau 1}U_{\tau 3}e^{-i\rho}\right)\right] \; ,$$
$$\sum_\alpha \left[U^*_{\alpha 2}\left(\dot{U}_{\alpha 3} + iU_{\alpha 3}\dot{\phi}_\alpha\right)\right] = -\frac{C_\kappa y^2_\tau}{16\pi^2} e^{i\sigma} \left[\zeta^{-1}_{23}{\rm Re}\left(U^*_{\tau 2}U_{\tau 3}e^{-i\sigma}\right) + i\zeta_{23}{\rm Im}\left(U^*_{\tau 2}U_{\tau 3}e^{-i\sigma}\right)\right] \; , \qquad (13$$

where $\zeta_{ij} \equiv (\kappa_i - \kappa_j)/(\kappa_i + \kappa_j)$ with $i \neq j$. One can see that those y^2_τ-associated terms only consist of the matrix elements $U_{\tau i}$ (for $i = 1, 2, 3$). If a parametrization of U assures $U_{\tau i}$ to be as simple as possible, the resultant RGEs of neutrino mixing angles and CP-violating phases will be very concise. We find that the FX parametrization advocated in Eq. (99) with

$$U = \begin{pmatrix} s_l s_\nu c + c_l c_\nu e^{-i\phi} & s_l c_\nu c - c_l s_\nu e^{-i\phi} & s_l s \\ c_l s_\nu c - s_l c_\nu e^{-i\phi} & c_l c_\nu c + s_l s_\nu e^{-i\phi} & c_l s \\ -s_\nu s & -c_\nu s & c \end{pmatrix}$$

accords with the above observation, while the "standard" parametrization in Eq. (98) does not. That is why the RGEs of neutrino mixing angles and CP-violating phases in the standard parametrization are rather complicated.

Here we take the FX form of U to derive the RGEs of neutrino mass and mixing parameters. Combining Eqs. (135), (136) and the FX form of U, we arrive at

$$\dot{\kappa}_1 = \frac{\kappa_1}{16\pi^2}\left(\alpha_\kappa + 2C_\kappa y^2_\tau s^2_\nu s^2\right) \; ,$$
$$\dot{\kappa}_2 = \frac{\kappa_2}{16\pi^2}\left(\alpha_\kappa + 2C_\kappa y^2_\tau c^2_\nu s^2\right) \; ,$$
$$\dot{\kappa}_3 = \frac{\kappa_3}{16\pi^2}\left(\alpha_\kappa + 2C_\kappa y^2_\tau c^2\right) \; , \qquad (137)$$

where $\alpha_\kappa \approx -3g^2_2 + 6y^2_t + \lambda$ (SM) or $\alpha_\kappa \approx -1.2g^2_1 - 6g^2_2 + 6y^2_t$ (MSSM); and

$$\dot{\theta}_l = \frac{C_\kappa y^2_\tau}{16\pi^2}\, c_\nu s_\nu c \left[\zeta^{-1}_{13}c_\rho c_{(\rho-\phi)} + \zeta_{13}s_\rho s_{(\rho-\phi)} - \zeta^{-1}_{23}c_\sigma c_{(\sigma-\phi)} - \zeta_{23}s_\sigma s_{(\sigma-\phi)}\right] \; ,$$
$$\dot{\theta}_\nu = \frac{C_\kappa y^2_\tau}{16\pi^2}\, c_\nu s_\nu \left[s^2\left(\zeta^{-1}_{12}c^2_{(\sigma-\rho)} + \zeta_{12}s^2_{(\sigma-\rho)}\right) + c^2\left(\zeta^{-1}_{13}c^2_\rho + \zeta_{13}s^2_\rho\right) - c^2\left(\zeta^{-1}_{23}c^2_\sigma + \zeta_{23}s^2_\sigma\right)\right] \; ,$$
$$\dot{\theta} = \frac{C_\kappa y^2_\tau}{16\pi^2}\, cs \left[s^2_\nu\left(\zeta^{-1}_{13}c^2_\rho + \zeta_{13}s^2_\rho\right) + c^2_\nu\left(\zeta^{-1}_{23}c^2_\sigma + \zeta_{23}s^2_\sigma\right)\right] \; ; \qquad (138)$$

as well as

$$\dot{\rho} = \frac{C_\kappa y^2_\tau}{16\pi^2}\left[\widehat{\zeta}_{12}c^2_\nu s^2 c_{(\sigma-\rho)}s_{(\sigma-\rho)} + \widehat{\zeta}_{13}\left(s^2_\nu s^2 - c^2\right)c_\rho s_\rho + \widehat{\zeta}_{23}c^2_\nu s^2 c_\sigma s_\sigma\right] \; ,$$
$$\dot{\sigma} = \frac{C_\kappa y^2_\tau}{16\pi^2}\left[\widehat{\zeta}_{12}s^2_\nu s^2 c_{(\sigma-\rho)}s_{(\sigma-\rho)} + \widehat{\zeta}_{13}s^2_\nu s^2 c_\rho s_\rho + \widehat{\zeta}_{23}\left(c^2_\nu s^2 - c^2\right)c_\sigma s_\sigma\right] \; ,$$
$$\dot{\phi} = \frac{C_\kappa y^2_\tau}{16\pi^2}\left[\left(c^2_l - s^2_l\right)c^{-1}_l s^{-1}_l c_\nu s_\nu c\left(\zeta^{-1}_{13}c_\rho s_{(\rho-\phi)} - \zeta_{13}s_\rho c_{(\rho-\phi)} - \zeta^{-1}_{23}c_\sigma s_{(\sigma-\phi)} + \zeta_{23}s_\sigma c_{(\sigma-\phi)}\right)\right.$$
$$\left. + \widehat{\zeta}_{12}s^2 c_{(\sigma-\rho)}s_{(\sigma-\rho)} + \widehat{\zeta}_{13}\left(s^2_\nu - c^2_\nu c^2\right)c_\rho s_\rho + \widehat{\zeta}_{23}\left(c^2_\nu - s^2_\nu c^2\right)c_\sigma s_\sigma\right] \; , \qquad (139)$$

where $\widehat{\zeta}_{ij} \equiv \zeta^{-1}_{ij} - \zeta_{ij} = 4\kappa_i\kappa_j / \left(\kappa^2_i - \kappa^2_j\right)$, $c_a \equiv \cos a$ and $s_a \equiv \sin a$ (for $a = \rho$, σ, $\sigma - \rho$, $\rho - \phi$ or $\sigma - \phi$).

Some discussions on the basic features of RGEs of three neutrino masses, three flavor mixing angles and three CP-violating phases are in order.

(a) The running behaviors of three neutrino masses m_i (or equivalently κ_i) are essentially identical and determined by α_κ, unless $\tan\beta$ is large enough in the MSSM to make the y^2_τ-associated term competitive with the α_κ term. In our phase convention, $\dot{\kappa}_i$ or $\dot{m}_i$ (for $i = 1, 2, 3$) are independent of the CP-violating phase ϕ.

(b) Among three neutrino mixing angles, only the derivative of θ_ν contains a term proportional to ζ^{-1}_{12}. Note that $\zeta^{-1}_{ij} = (m_i + m_j)^2/\Delta m^2_{ij}$ with $\Delta m^2_{ij} \equiv m^2_i - m^2_j$ holds. Current solar and atmospheric neutrino oscillation data yield $\Delta m^2_{21} \approx 7.7 \times 10^{-5}$ eV^2 and $\left|\Delta m^2_{32}\right| \approx \left|\Delta m^2_{31}\right| \approx 2.4 \times 10^{-3}$ eV^2. So θ_ν is in general more sensitive to radiative corrections than θ_l and θ. The evolution of θ_ν can be suppressed through the fine-tuning of $(\sigma - \rho)$. The smallest neutrino mixing angle θ_l may get radiative corrections even if its initial value is zero, and thus it can be radiatively generated from other neutrino mixing angles and CP-violating phases.

(c) The running behavior of ϕ is quite different from those of ρ and σ, because it includes a peculiar term proportional to s^{-1}_l. This term, which dominates $\dot{\phi}$ when θ_l is sufficiently small, becomes divergent in the limit $\theta_l \to 0$. Indeed, ϕ is not well-defined if θ_l is exactly vanishing. But both θ_l and ϕ can be radiatively generated. We may require that $\dot{\phi}$ remain finite when θ_l approaches zero, implying that the following necessary condition can be extracted from the expression of $\dot{\phi}$ in Eq. (139):

$$\zeta^{-1}_{13} c_\rho s_{(\rho-\phi)} - \zeta_{13} s_\rho c_{(\rho-\phi)} - \zeta^{-1}_{23} c_\sigma s_{(\sigma-\phi)} + \zeta_{23} s_\sigma c_{(\sigma-\phi)} = 0 \; . \tag{140}$$

Note that the initial value of θ_l, if it is exactly zero or extremely small, may immediately drive ϕ to its *quasi-fixed point*. In this case Eq. (140) can be used to understand the relationship between ϕ and two Majorana phases ρ and σ at the quasi-fixed point.

(d) The running behaviors of ρ and σ are relatively mild in comparison with that of ϕ. A remarkable feature of $\dot{\rho}$ and $\dot{\sigma}$ is that they will vanish, if both ρ and σ are initially vanishing. This observation indicates that ρ and σ cannot simultaneously be generated from ϕ via the RGEs.

6 How to Generate Neutrino Masses?

Neutrinos are assumed or required to be massless in the SM, just because the structure of the SM itself is too simple to accommodate massive neutrinos.

- Two fundamentals of the SM are the $SU(2)_{\rm L} \times U(1)_{\rm Y}$ gauge symmetry and the Lorentz invariance. Both of them are mandatory to guarantee that the SM is a consistent quantum field theory.
- The particle content of the SM is rather economical. There are no right-handed neutrinos in the SM, so a Dirac neutrino mass term is not allowed. There is only one Higgs doublet, so a gauge-invariant Majorana mass term is forbidden.
- The SM is a renormalizable quantum field theory. Hence an effective dimension-5 operator, which may give each neutrino a Majorana mass, is absent.

In other words, the SM accidently possesses the $(B - L)$ symmetry which assures three known neutrinos to be exactly massless.

But today's experiments have convincingly indicated the existence of neutrino oscillations. This quantum phenomenon can appear if and only if neutrinos are massive and lepton flavors are mixed, and thus it is a kind of new physics beyond the SM. To generate non-zero but tiny neutrino masses, one or more of the above-mentioned constraints on the SM must be abandoned or relaxed. It is intolerable to abandon the gauge symmetry and Lorentz invariance; otherwise, one would be led astray. Given

the framework of the SM as a consistent field theory, its particle content can be modified and (or) its renormalizability can be abandoned to accommodate massive neutrinos. There are several ways to this goal.

6.1 Relaxing the renormalizability

In 1979, Weinberg extended the SM by introducing some higher-dimension operators in terms of the fields of the SM itself:

$$\mathcal{L}_{\rm eff} = \mathcal{L}_{\rm SM} + \frac{\mathcal{L}_{\rm d=5}}{\Lambda} + \frac{\mathcal{L}_{\rm d=6}}{\Lambda^2} + \cdots \; , \tag{141}$$

where Λ denotes the cut-off scale of this effective theory. Within such a framework, the lowest-dimension operator that violates the lepton number (L) is the unique dimension-5 operator $HHLL/\Lambda$. After spontaneous gauge symmetry breaking, this Weinberg operator yields $m_i \sim \langle H \rangle^2/\Lambda$ for neutrino masses, which can be sufficiently small (≤ 1 eV) if Λ is not far away from the scale of grand unified theories ($\Lambda \sim 10^{13}$ GeV for $\langle H \rangle \sim 10^2$ GeV). In this sense we argue that neutrino masses can serve as a low-energy window onto new physics at superhigh energies.

6.2 A pure Dirac neutrino mass term?

Given three right-handed neutrinos, the gauge-invariant and lepton-number-conserving mass terms of charged leptons and neutrinos are

$$-\mathcal{L}_{\rm lepton} = \overline{\ell_{\rm L}} Y_l H E_{\rm R} + \overline{\ell_{\rm L}} Y_\nu \tilde{H} N_{\rm R} + {\rm h.c.} \; , \tag{142}$$

where $\tilde{H} \equiv i\sigma_2 H^*$ is defined and $\ell_{\rm L}$ denotes the left-handed lepton doublet. After spontaneous gauge symmetry breaking, we arrive at the charged-lepton mass matrix $M_l = Y_l v/\sqrt{2}$ and the Dirac neutrino mass matrix $M_\nu = Y_\nu v/\sqrt{2}$ with $v \simeq 246$ GeV. In this case, the smallness of three neutrino masses m_i (for $i = 1, 2, 3$) is attributed to the smallness of three eigenvalues of Y_ν (denoted as y_i for $i = 1, 2, 3$). Then we encounter a transparent hierarchy problem: $y_i/y_e = m_i/m_e \leq 0.5\,{\rm eV}/0.5\,{\rm MeV} \sim 10^{-6}$. Why is y_i so small? There is no explanation at all in this Dirac-mass picture.

A speculative way out is to invoke extra dimensions; namely, the smallness of Dirac neutrino masses is ascribed to the assumption that three right-handed neutrinos have access to one or more extra spatial dimensions. The idea is simply to confine the SM particles onto a brane and to allow $N_{\rm R}$ to travel in the bulk. For example, the wave-function of $N_{\rm R}$ spreads out over the extra dimension y, giving rise to a suppressed Yukawa interaction at $y = 0$ (i.e., the location of the brane):

$$\left[\overline{\ell_{\rm L}} Y_\nu \tilde{H} N_{\rm R}\right]_{y=0} \sim \frac{1}{\sqrt{L}} \left[\overline{\ell_{\rm L}} Y_\nu \tilde{H} N_{\rm R}\right]_{y=L} \; . \tag{143}$$

The magnitude of $1/\sqrt{L}$ is measured by $\Lambda/\Lambda_{\rm Planck}$, and thus it can naturally be small for an effective theory far below the Planck scale.

6.3 Seesaw mechanisms

This approach works at the tree level and reflects the essential spirit of seesaw mechanisms — tiny masses of three known neutrinos are attributed to the existence of heavy degrees of freedom and lepton number violation.

– Type-I seesaw — three heavy right-handed neutrinos are added into the SM and the lepton number is violated by their Majorana mass term:

$$-\mathcal{L}_{\rm lepton} = \overline{\ell_{\rm L}} Y_l H E_{\rm R} + \overline{\ell_{\rm L}} Y_\nu \tilde{H} N_{\rm R} + \frac{1}{2} \overline{N_{\rm R}^{\rm c}} M_{\rm R} N_{\rm R} + {\rm h.c.} \; , \tag{144}$$

where $M_{\rm R}$ is the Majorana mass matrix.

– Type-II seesaw — one heavy Higgs triplet is added into the SM and the lepton number is violated by its interactions with both the lepton doublet and the Higgs doublet:

$$-\mathcal{L}_{\rm lepton} = \overline{\ell_{\rm L}} Y_l H E_{\rm R} + \frac{1}{2}\overline{\ell_{\rm L}} Y_\Delta \Delta i\sigma_2 \ell^{\rm c}_{\rm L} - \lambda_\Delta M_\Delta H^T i\sigma_2 \Delta H + {\rm h.c.} \; , \tag{145}$$

where

$$\Delta \equiv \begin{pmatrix} \Delta^- & -\sqrt{2}\,\Delta^0 \\ \sqrt{2}\,\Delta^{--} & -\Delta^- \end{pmatrix} \tag{146}$$

denotes the $SU(2)_{\rm L}$ Higgs triplet.

– Type-III seesaw — three heavy triplet fermions are added into the SM and the lepton number is violated by their Majorana mass term:

$$-\mathcal{L}_{\rm lepton} = \overline{\ell_{\rm L}} Y_l H E_{\rm R} + \overline{\ell_{\rm L}} \sqrt{2} Y_\Sigma \Sigma^{\rm c} \tilde{H} + \frac{1}{2}{\rm Tr}\left(\overline{\Sigma} M_\Sigma \Sigma^{\rm c}\right) + {\rm h.c.} \; , \tag{147}$$

where

$$\Sigma = \begin{pmatrix} \Sigma^0/\sqrt{2} & \Sigma^+ \\ \Sigma^- & -\Sigma^0/\sqrt{2} \end{pmatrix} \tag{148}$$

denotes the $SU(2)_{\rm L}$ fermion triplet.

Of course, there are a number of variations or combinations of these three typical seesaw mechanisms in the literature.

For each of the above seesaw pictures, one may arrive at the unique dimension-5 Weinberg operator of neutrino masses after integrating out the corresponding heavy degrees of freedom:

$$\frac{\mathcal{L}_{\rm d=5}}{\Lambda} = \begin{cases} \frac{1}{2}\left(Y_\nu M_{\rm R}^{-1} Y_\nu^T\right)_{\alpha\beta} \overline{\ell_{\alpha{\rm L}}} \tilde{H} \tilde{H}^T \ell^{\rm c}_{\beta{\rm L}} + {\rm h.c.} \\ -\frac{\lambda_\Delta}{M_\Delta}\left(Y_\Delta\right)_{\alpha\beta} \overline{\ell_{\alpha{\rm L}}} \tilde{H} \tilde{H}^T \ell^{\rm c}_{\beta{\rm L}} + {\rm h.c.} \\ \frac{1}{2}\left(Y_\Sigma M_\Sigma^{-1} Y_\Sigma^T\right)_{\alpha\beta} \overline{\ell_{\alpha{\rm L}}} \tilde{H} \tilde{H}^T \ell^{\rm c}_{\beta{\rm L}} + {\rm h.c.} \end{cases}$$

corresponding to type-I, type-II and type-III seesaws. After spontaneous gauge symmetry breaking, $\tilde{H}$ achieves its vacuum expectation value $\langle \tilde{H} \rangle = v/\sqrt{2}$ with $v \simeq 246$ GeV. Then we are left with the effective Majorana neutrino mass term for three known neutrinos,

$$-\mathcal{L}_{\rm mass} = \frac{1}{2}\overline{\nu_{\rm L}} M_\nu \nu^{\rm c}_{\rm L} + {\rm h.c.} \; , \tag{149}$$

where the Majorana mass matrix M_ν is given by

$$M_\nu = \begin{cases} -\frac{1}{2} Y_\nu \frac{v^2}{M_{\rm R}} Y_\nu^T & ({\rm Type\ I}) \; , \\ \lambda_\Delta Y_\Delta \frac{v^2}{M_\Delta} & ({\rm Type\ II}) \; , \\ -\frac{1}{2} Y_\Sigma \frac{v^2}{M_\Sigma} Y_\Sigma^T & ({\rm Type\ III}) \; . \end{cases} \tag{150}$$

It becomes obvious that the smallness of M_ν can be attributed to the largeness of $M_{\rm R}$, M_Δ or M_Σ in the seesaw mechanism.

6.4 Radiative origin of neutrino masses

In a seminal paper published in 1972, Weinberg pointed out that "in theories with spontaneously broken gauge symmetries, various masses or mass differences may vanish in zeroth order as a consequence of the representation content of the fields appearing in the Lagrangian. These masses or mass differences can then be calculated as finite higher-order effects." Such a mechanism may allow us to slightly go beyond the SM and radiatively generate tiny neutrino masses. A typical example is the well-known Zee model,

$$-\mathcal{L}_{\rm lepton} = \overline{\ell_{\rm L}} Y_l H E_{\rm R} + \overline{\ell_{\rm L}} Y_S S^- i\sigma_2 l^{\rm c}_{\rm L} + \tilde{\Phi}^T F S^+ i\sigma_2 \tilde{H} + {\rm h.c.} \; , \tag{151}$$

where $S^\pm$ are charged $SU(2)_{\rm L}$ singlet scalars, Φ denotes a new $SU(2)_{\rm L}$ doublet scalar which has the same quantum number as the SM Higgs doublet H, Y_S is an anti-symmetric matrix, and F represents a mass. Without loss of generality, we choose the basis of $M_l = Y_l \langle H \rangle = {\rm Diag}\{m_e, m_\mu, m_\tau\}$. In this model neutrinos are massless at the tree level, but their masses can radiatively be generated via the one-loop corrections. Given $M_S \gg M_H \sim M_\Phi \sim F$ and $\langle \Phi \rangle \sim \langle H \rangle$, the elements of the effective mass matrix of three light Majorana neutrinos are

$$(M_\nu)_{\alpha\beta} \; \sim \; \frac{M_H}{16\pi^2} \cdot \frac{m^2_\alpha - m^2_\beta}{M^2_S} \, (Y_S)_{\alpha\beta} \; , \tag{152}$$

where α and β run over e, μ and τ. The smallness of M_ν is therefore ascribed to the smallness of Y_S and $(m^2_\alpha - m^2_\beta)/M^2_S$. Although the original version of the Zee model is disfavored by current experimental data on neutrino oscillations, its extensions or variations at the one-loop or two-loop level can survive.

7 On the Scales of Seesaw Mechanisms

As we have seen, the key point of a seesaw mechanism is to ascribe the smallness of neutrino masses to the existence of some new degrees of freedom heavier than the Fermi scale $v \simeq 246$ GeV, such as heavy Majorana neutrinos or heavy Higgs bosons. The energy scale where a seesaw mechanism works is crucial, because it is relevant to whether this mechanism is theoretically natural and experimentally testable. Between Fermi and Planck scales, there might exist two other fundamental scales: one is the scale of a grand unified theory (GUT) at which strong, weak and electromagnetic forces can be unified, and the other is the TeV scale at which the unnatural gauge hierarchy problem of the SM can be solved or at least softened by a kind of new physics.

7.1 How about a very low seesaw scale?

In reality, however, there is no direct evidence for a high or extremely high seesaw scale. Hence eV-, keV-, MeV- and GeV-scale seesaws are all possible, at least in principle, and they are technically natural in the sense that their lepton-number-violating mass terms are naturally small according to 't Hooft's naturalness criterion — "At any energy scale μ, a set of parameters $\alpha_i(\mu)$ describing a system can be small, if and only if, in the limit $\alpha_i(\mu) \to 0$ for each of these parameters, the system exhibits an enhanced symmetry." But there are several potential problems associated with low-scale seesaws: (a) a low-scale seesaw does not give any obvious connection to a theoretically well-justified fundamental physical scale (such as the Fermi scale, the TeV scale, the GUT scale or the Planck scale); (b) the neutrino Yukawa couplings in a low-scale seesaw model turn out to be tiny, giving no actual explanation of why the masses of three known neutrinos are so small; and (c) in general, a very low seesaw scale does not allow the "canonical" thermal leptogenesis mechanism to work.

7.2 Seesaw-induced hierarchy problem

Many theorists argue that the conventional seesaw scenarios are natural because their scales (i.e., the masses of heavy degrees of freedom) are close to the GUT scale. This argument is reasonable on the one

hand, but it reflects the drawbacks of the conventional seesaw models on the other hand. In other words, the conventional seesaw models have no direct experimental testability and involve a potential hierarchy problem. The latter is usually spoke of when two largely different energy scales exist in a model, but there is no symmetry to stabilize the low-scale physics suffering from large corrections coming from the high-scale physics.

Such a seesaw-induced fine-tuning problem means that the SM Higgs mass is very sensitive to quantum corrections from the heavy degrees of freedom in a seesaw mechanism. For example,

$$\delta M_H^2 = \begin{cases} -\dfrac{y_i^2}{8\pi^2}\left(\Lambda^2 + M_i^2 \ln \dfrac{M_i^2}{\Lambda^2}\right) & \text{(I)} \\ \dfrac{3}{16\pi^2}\left[\lambda_3\left(\Lambda^2 + M_\Delta^2 \ln \dfrac{M_\Delta^2}{\Lambda^2}\right) + 4\lambda_\Delta^2 M_\Delta^2 \ln \dfrac{M_\Delta^2}{\Lambda^2}\right] & \text{(II)} \\ -\dfrac{3y_i^2}{8\pi^2}\left(\Lambda^2 + M_i^2 \ln \dfrac{M_i^2}{\Lambda^2}\right) & \text{(III)} \end{cases}$$

in three typical seesaw scenarios, where Λ is the regulator cut-off, y_i and M_i (for $i = 1, 2, 3$) stand respectively for the eigenvalues of Y_ν (or Y_Σ) and $M_{\rm R}$ (or M_Σ), and the contributions proportional to v^2 and M_H^2 have been omitted. The above results show a quadratic sensitivity to the new scale which is characteristic of the seesaw model, implying that a high degree of fine-tuning would be necessary to accommodate the experimental data on M_H if the seesaw scale is much larger than v (or the Yukawa couplings are not extremely fine-tuned in type-I and type-III seesaws). Taking the type-I seesaw scenario for illustration, we assume $\Lambda \sim M_i$ and require $|\delta M_H^2| \leq 0.1~{\rm TeV}^2$. Then the above equation leads us to the following rough estimate:

$$M_i \sim \left[\frac{(2\pi v)^2 |\delta M_H^2|}{m_i}\right]^{1/3} \leq 10^7 {\rm GeV} \left[\frac{0.2~{\rm eV}}{m_i}\right]^{1/3} \left[\frac{|\delta M_H^2|}{0.1~{\rm TeV}^2}\right]^{1/3} . \tag{153}$$

This naive result indicates that a hierarchy problem will arise if the masses of heavy Majorana neutrinos are larger than about 10^7 GeV in the type-I seesaw scheme. Because of $m_i \sim y_i^2 v^2/(2M_i)$, the bound $M_i \leq 10^7$ GeV implies $y_i \sim \sqrt{2m_i M_i}/v \leq 2.6 \times 10^{-4}$ for $m_i \sim 0.2$ eV. Such a small magnitude of y_i seems to be a bit unnatural in the sense that the conventional seesaw idea attributes the smallness of m_i to the largeness of M_i other than the smallness of y_i.

There are two possible ways out of this impasse: one is to appeal for the supersymmetry, and the other is to lower the seesaw scale. We shall follow the second way to discuss the TeV seesaw mechanisms which do not suffer from the above-mentioned hierarchy problem.

7.3 Why are the TeV seesaws interesting?

There are several reasons for people to expect some new physics at the TeV scale. This kind of new physics should be able to stabilize the Higgs-boson mass and hence the electroweak scale; in other words, it should be able to solve or soften the unnatural gauge hierarchy problem. It has also been argued that the weakly-interacting particle candidates for dark matter should weigh about one TeV or less. If the TeV scale is really a fundamental scale, may we argue that the TeV seesaws are natural? Indeed, we are reasonably motivated to speculate that possible new physics existing at the TeV scale and responsible for the electroweak symmetry breaking might also be responsible for the origin of neutrino masses. It is interesting and meaningful in this sense to investigate and balance the “naturalness" and “testability" of TeV seesaws at the energy frontier set by the LHC.

As a big bonus of the conventional (type-I) seesaw mechanism, the thermal leptogenesis mechanism provides us with an elegant dynamic picture to interpret the cosmological matter-antimatter asymmetry characterized by the observed ratio of baryon number density to photon number density, $\eta_B \equiv$

$n_B/n_\gamma = (6.1 \pm 0.2) \times 10^{10}$. When heavy Majorana neutrino masses are down to the TeV scale, the Yukawa couplings should be reduced by more than six orders of magnitude so as to generate tiny masses for three known neutrinos via the type-I seesaw and satisfy the out-of-equilibrium condition, but the CP-violating asymmetries of heavy Majorana neutrino decays can still be enhanced by the resonant effects in order to account for η_B. This "resonant leptogenesis" scenario might work in a specific TeV seesaw model.

Is there a TeV Noah's Ark which can naturally and simultaneously accommodate the seesaw idea, the leptogenesis picture and the collider signatures? We are most likely not so lucky and should not be too ambitious at present. In the following we shall concentrate on the TeV seesaws themselves and their possible collider signatures and low-energy consequences.

8 TeV Seesaws: Natural and Testable?

The neutrino mass terms in three typical seesaw mechanisms have been given before. Without loss of generality, we choose the basis in which the mass eigenstates of three charged leptons are identified with their flavor eigenstates.

8.1 Type-I seesaw

Given $M_{\rm D} = Y_\nu v/\sqrt{2}$, the approximate type-I seesaw formula in Eq. (150) can be rewritten as $M_\nu = -M_{\rm D} M_{\rm R}^{-1} M_{\rm D}^T$. Note that the 3×3 light neutrino mixing matrix V is not exactly unitary in this seesaw scheme, and its deviation from unitarity is of $\mathcal{O}(M_{\rm D}^2/M_{\rm R}^2)$. Let us consider two interesting possibilities. (1) $M_{\rm D} \sim \mathcal{O}(10^2)$ GeV and $M_{\rm R} \sim \mathcal{O}(10^{15})$ GeV to get $M_\nu \sim \mathcal{O}(10^{-2})$ eV. In this conventional and *natural* case, $M_{\rm D}/M_{\rm R} \sim \mathcal{O}(10^{-13})$ holds. Hence the non-unitarity of V is only at the $\mathcal{O}(10^{-26})$ level, too small to be observed. (2) $M_{\rm D} \sim \mathcal{O}(10^2)$ GeV and $M_{\rm R} \sim \mathcal{O}(10^3)$ GeV to get $M_\nu \sim \mathcal{O}(10^{-2})$ eV. In this *unnatural* case, a significant "structural cancellation" has to be imposed on the textures of $M_{\rm D}$ and $M_{\rm R}$. Because of $M_{\rm D}/M_{\rm R} \sim \mathcal{O}(0.1)$, the non-unitarity of V can reach the percent level and may lead to observable effects.

Now we discuss how to realize the above "structural cancellation" for the type-I seesaw mechanism at the TeV scale. For the sake of simplicity, we take the basis of $M_{\rm R} = {\rm Diag}\{M_1, M_2, M_3\}$ for three heavy Majorana neutrinos (N_1, N_2, N_3). It is well known that M_ν vanishes if

$$M_{\rm D} = m \begin{pmatrix} y_1 & y_2 & y_3 \\ \alpha y_1 & \alpha y_2 & \alpha y_3 \\ \beta y_1 & \beta y_2 & \beta y_3 \end{pmatrix} \; , \quad \sum_{i=1}^{3} \frac{y_i^2}{M_i} = 0 \tag{154}$$

simultaneously hold. Tiny neutrino masses can be generated from tiny corrections to the texture of $M_{\rm D}$ in Eq. (154). For example, $M'_{\rm D} = M_{\rm D} - \epsilon X_{\rm D}$ with $M_{\rm D}$ given above and ϵ being a small dimensionless parameter (i.e., $|\epsilon| \ll 1$) yields

$$M'_\nu = -M'_{\rm D} M_{\rm R}^{-1} M'^T_{\rm D} \simeq \epsilon \left(M_{\rm D} M_{\rm R}^{-1} X_{\rm D}^T + X_{\rm D} M_{\rm R}^{-1} M_{\rm D}^T \right) \; , \tag{155}$$

from which $M'_\nu \sim \mathcal{O}(10^{-2})$ eV can be obtained by adjusting the size of ϵ.

A lot of attention has recently been paid to a viable type-I seesaw model and its collider signatures at the TeV scale. At least the following lessons can be learnt:

- Two necessary conditions must be satisfied in order to test a type-I seesaw model at the LHC: (a) M_i are of $\mathcal{O}(1)$ TeV or smaller; and (b) the strength of light-heavy neutrino mixing (i.e., $M_{\rm D}/M_{\rm R}$) is large enough. Otherwise, it would be impossible to produce and detect N_i at the LHC.
- The collider signatures of N_i are essentially decoupled from the mass and mixing parameters of three light neutrinos ν_i. For instance, the small parameter ϵ in Eq. (155) has nothing to do with the ratio $M_{\rm D}/M_{\rm R}$.

– The non-unitarity of V might lead to some observable effects in neutrino oscillations and other lepton-flavor-violating or lepton-number-violating processes, if $M_{\rm D}/M_{\rm R} \leq \mathcal{O}(0.1)$ holds.
– The clean LHC signatures of heavy Majorana neutrinos are the $\Delta L = 2$ like-sign dilepton events, such as $pp \to W^{*\pm} W^{*\pm} \to \mu^\pm \mu^\pm jj$ and $pp \to W^{*\pm} \to \mu^\pm N_i \to \mu^\pm \mu^\pm jj$ (a dominant channel due to the resonant production of N_i).

Some instructive and comprehensive analyses of possible LHC events for a single heavy Majorana neutrino have recently been done, but they only serve for illustration because such a simplified type-I seesaw scenario is actually unrealistic.

8.2 Type-II seesaw

The type-II seesaw formula $M_\nu = Y_\Delta v_\Delta = \lambda_\Delta Y_\Delta v^2/M_\Delta$ has been given in Eq. (150). Note that the last term of Eq. (145) violates both L and $B - L$, and thus the smallness of λ_Δ is naturally allowed according to 't Hooft's naturalness criterion (i.e., setting $\lambda_\Delta = 0$ will increase the symmetry of $\mathcal{L}_{\rm lepton}$). Given $M_\Delta \sim \mathcal{O}(1)$ TeV, for example, this seesaw mechanism works to generate $M_\nu \sim \mathcal{O}(10^{-2})$ eV provided $\lambda_\Delta Y_\Delta \sim \mathcal{O}(10^{-12})$ holds. The neutrino mixing matrix V is exactly unitary in the type-II seesaw mechanism, simply because the heavy degrees of freedom do not mix with the light ones.

There are totally seven physical Higgs bosons in the type-II seesaw scheme: doubly-charged H^{++} and H^{--}, singly-charged H^+ and H^-, neutral A^0 (CP-odd), and neutral h^0 and H^0 (CP-even), where h^0 is the SM-like Higgs boson. Except for $M^2_{h^0}$, we get a quasi-degenerate mass spectrum for other scalars: $M^2_{H^{\pm\pm}} = M^2_\Delta \approx M^2_{H^0} \approx M^2_{H^\pm} \approx M^2_{A^0}$. As a consequence, the decay channels $H^{\pm\pm} \to W^\pm H^\pm$ and $H^{\pm\pm} \to H^\pm H^\pm$ are kinematically forbidden. The production of $H^{\pm\pm}$ at the LHC is mainly through $q\bar{q} \to \gamma^*, Z^* \to H^{++} H^{--}$ and $q\bar{q}' \to W^* \to H^{\pm\pm} H^\mp$ processes, which do not rely on the small Yukawa couplings.

The typical collider signatures in this seesaw scenario are the lepton-number-violating $H^{\pm\pm} \to l^\pm_\alpha l^\pm_\beta$ decays as well as $H^+ \to l^+_\alpha \overline{\nu}$ and $H^- \to l^-_\alpha \nu$ decays. Their branching ratios

$$\begin{aligned} \mathcal{B}(H^{\pm\pm} \to l^\pm_\alpha l^\pm_\beta) &= \frac{|(M_\nu)_{\alpha\beta}|^2 \left(2 - \delta_{\alpha\beta}\right)}{\sum\limits_{\rho,\sigma} |(M_\nu)_{\rho\sigma}|^2} \; , \\ \mathcal{B}(H^+ \to l^+_\alpha \overline{\nu}) &= \frac{\sum\limits_\beta |(M_\nu)_{\alpha\beta}|^2}{\sum\limits_{\rho,\sigma} |(M_\nu)_{\rho\sigma}|^2} \end{aligned} \tag{156}$$

are closely related to the masses, flavor mixing angles and CP-violating phases of three light neutrinos, because $M_\nu = V \widehat{M}_\nu V^T$ with $\widehat{M}_\nu = {\rm Diag}\{m_1, m_2, m_3\}$ holds. Some detailed analyses of such decay modes together with the LHC signatures of $H^{\pm\pm}$ and $H^\pm$ bosons have been done in the literature.

It is worth pointing out that the following dimension-6 operator can easily be derived from the type-II seesaw mechanism,

$$\frac{\mathcal{L}_{\rm d=6}}{\Lambda^2} = -\frac{(Y_\Delta)_{\alpha\beta} (Y_\Delta)^\dagger_{\rho\sigma}}{4 M^2_\Delta} (\overline{\ell_{\alpha {\rm L}}} \gamma^\mu \ell_{\sigma {\rm L}}) (\overline{\ell_{\beta {\rm L}}} \gamma_\mu \ell_{\rho {\rm L}}) \; , \tag{157}$$

which has two immediate low-energy effects: the non-standard interactions of neutrinos and the lepton-flavor-violating interactions of charged leptons. An analysis of such effects provides us with some preliminary information:

– The magnitudes of non-standard interactions of neutrinos and the widths of lepton-flavor-violating tree-level decays of charged leptons are both dependent on neutrino masses m_i and flavor-mixing and CP-violating parameters of V.

– For a long-baseline neutrino oscillation experiment, the neutrino beam encounters the earth matter and the electron-type non-standard interaction contributes to the matter potential.
– At a neutrino factory, the lepton-flavor-violating processes $\mu^- \to e^- \nu_e \overline{\nu}_\mu$ and $\mu^+ \to e^+ \overline{\nu}_e \nu_\mu$ could cause some wrong-sign muons at a near detector.

Current experimental constraints tell us that such low-energy effects are very small, but they might be experimentally accessible in the future precision measurements.

8.3 Type-(I+II) seesaw

The type-(I+II) seesaw mechanism can be achieved by combining the neutrino mass terms in Eqs. (144) and (145). After spontaneous gauge symmetry breaking, we are left with the overall neutrino mass term

$$-\mathcal{L}_{\rm mass} = \frac{1}{2} \overline{(\nu_{\rm L} N^{\rm c}_{\rm R})} \begin{pmatrix} M_{\rm L} & M_{\rm D} \\ M^T_{\rm D} & M_{\rm R} \end{pmatrix} \begin{pmatrix} \nu^{\rm c}_{\rm L} \\ N_{\rm R} \end{pmatrix} + {\rm h.c.} \; , \tag{158}$$

where $M_{\rm D} = Y_\nu v/\sqrt{2}$ and $M_{\rm L} = Y_\Delta v_\Delta$ with $\langle H \rangle \equiv v/\sqrt{2}$ and $\langle \Delta \rangle \equiv v_\Delta$ corresponding to the vacuum expectation values of the neutral components of the Higgs doublet H and the Higgs triplet Δ. The 6×6 mass matrix in Eq. (158) is symmetric and can be diagonalized by the unitary transformation done in Eq. (28); i.e.,

$$\begin{pmatrix} V & R \\ S & U \end{pmatrix}^\dagger \begin{pmatrix} M_{\rm L} & M_{\rm D} \\ M^T_{\rm D} & M_{\rm R} \end{pmatrix} \begin{pmatrix} V & R \\ S & U \end{pmatrix}^* = \begin{pmatrix} \widehat{M}_\nu & \mathbf{0} \\ \mathbf{0} & \widehat{M}_N \end{pmatrix} \; , \tag{159}$$

where $\widehat{M}_\nu = {\rm Diag}\{m_1, m_2, m_3\}$ and $\widehat{M}_N = {\rm Diag}\{M_1, M_2, M_3\}$. Needless to say, $V^\dagger V + S^\dagger S = V V^\dagger + R R^\dagger = \mathbf{1}$ holds as a consequence of the unitarity of this transformation. Hence V, the flavor mixing matrix of light Majorana neutrinos, must be non-unitary if R and S are non-zero.

In the leading-order approximation, the type-(I+II) seesaw formula reads as

$$M_\nu \approx M_{\rm L} - M_{\rm D} M^{-1}_{\rm R} M^T_{\rm D} \; . \tag{160}$$

Hence type-I and type-II seesaws can be regarded as two extreme cases of the type-(I+II) seesaw. Note that two mass terms in Eq. (160) are possibly comparable in magnitude. If both of them are small, their contributions to M_ν may have significant interference effects which make it practically impossible to distinguish between type-II and type-(I+II) seesaws; but if both of them are large, their contributions to M_ν must be destructive. The latter case unnaturally requires a significant cancellation between two big quantities in order to obtain a small quantity, but it is interesting in the sense that it may give rise to possibly observable collider signatures of heavy Majorana neutrinos.

Let me briefly describe a particular type-(I+II) seesaw model and comment on its possible LHC signatures. First, we assume that both M_i and M_Δ are of $\mathcal{O}(1)$ TeV. Then the production of $H^{\pm\pm}$ and $H^\pm$ bosons at the LHC is guaranteed, and their lepton-number-violating signatures will probe the Higgs triplet sector of the type-(I+II) seesaw mechanism. On the other hand, $\mathcal{O}(M_{\rm D}/M_{\rm R}) \leq \mathcal{O}(0.1)$ is possible as a result of $\mathcal{O}(M_{\rm R}) \sim \mathcal{O}(1)$ TeV and $\mathcal{O}(M_{\rm D}) \leq \mathcal{O}(v)$, such that appreciable signatures of N_i can be achieved at the LHC. Second, the small mass scale of M_ν implies that the relation $\mathcal{O}(M_{\rm L}) \sim \mathcal{O}(M_{\rm D} M^{-1}_{\rm R} M^T_{\rm D})$ must hold. In other words, it is the significant but incomplete cancellation between $M_{\rm L}$ and $M_{\rm D} M^{-1}_{\rm R} M^T_{\rm D}$ terms that results in the non-vanishing but tiny masses for three light neutrinos. We admit that dangerous radiative corrections to two mass terms of M_ν require a delicate fine-tuning of the cancellation at the loop level. But this scenario allows us to reconstruct $M_{\rm L}$ via the excellent approximation $M_{\rm L} = V \widehat{M}_\nu V^T + R \widehat{M}_N R^T \approx R \widehat{M}_N R^T$, such that the elements of the Yukawa coupling matrix Y_Δ read as follows:

$$(Y_\Delta)_{\alpha\beta} \; = \; \frac{(M_{\rm L})_{\alpha\beta}}{v_\Delta} \approx \sum_{i=1}^{3} \frac{R_{\alpha i} R_{\beta i} M_i}{v_\Delta} \; , \tag{161}$$

where the subscripts α and β run over e, μ and τ. This result implies that the leptonic decays of $H^{\pm\pm}$ and $H^{\pm}$ bosons depend on both R and M_i, which actually determine the production and decays of N_i. Thus we have established an interesting correlation between the singly- or doubly-charged Higgs bosons and the heavy Majorana neutrinos. To observe the correlative signatures of $H^{\pm}$, $H^{\pm\pm}$ and N_i at the LHC will serve for a direct test of this type-(I+II) seesaw model.

8.4 Type-III seesaw

The lepton mass terms in the type-III seesaw scheme have already been given in Eq. (147). After spontaneous gauge symmetry breaking, we are left with

$$-\mathcal{L}_{\rm mass} = \frac{1}{2}\, \overline{\left(\nu_{\rm L}\ \Sigma^0\right)} \begin{pmatrix} \mathbf{0} & M_{\rm D} \\ M_{\rm D}^T & M_\Sigma \end{pmatrix} \begin{pmatrix} \nu_{\rm L}^{\rm c} \\ \Sigma^{0{\rm c}} \end{pmatrix} + {\rm h.c.} \; ,$$
$$-\mathcal{L}'_{\rm mass} = \overline{\left(e_{\rm L}\ \Psi_{\rm L}\right)} \begin{pmatrix} M_l & \sqrt{2} M_{\rm D} \\ \mathbf{0} & M_\Sigma \end{pmatrix} \begin{pmatrix} E_{\rm R} \\ \Psi_{\rm R} \end{pmatrix} + {\rm h.c.} \; , \tag{162}$$

respectively, for neutral and charged fermions, where $M_l = Y_l v/\sqrt{2}$, $M_{\rm D} = Y_\Sigma v/\sqrt{2}$, and $\Psi = \Sigma^- + \Sigma^{+{\rm c}}$. The symmetric 6×6 neutrino mass matrix can be diagonalized by the following unitary transformation:

$$\begin{pmatrix} V & R \\ S & U \end{pmatrix}^\dagger \begin{pmatrix} \mathbf{0} & M_{\rm D} \\ M_{\rm D}^T & M_\Sigma \end{pmatrix} \begin{pmatrix} V & R \\ S & U \end{pmatrix}^* = \begin{pmatrix} \widehat{M}_\nu & \mathbf{0} \\ \mathbf{0} & \widehat{M}_\Sigma \end{pmatrix} \; , \tag{163}$$

where $\widehat{M}_\nu = {\rm Diag}\{m_1, m_2, m_3\}$ and $\widehat{M}_\Sigma = {\rm Diag}\{M_1, M_2, M_3\}$. In the leading-order approximation, this diagonalization yields the type-III seesaw formula $M_\nu = -M_{\rm D} M_\Sigma^{-1} M_{\rm D}^T$, which is equivalent to the one derived from the effective dimension-5 operator in Eq. (150). Let us use one sentence to comment on the similarities and differences between type-I and type-III seesaw mechanisms: the non-unitarity of the 3×3 neutrino mixing matrix V has appeared in both cases, although the modified couplings between the Z^0 boson and three light neutrinos differ and the non-unitary flavor mixing is also present in the couplings between the Z^0 boson and three charged leptons in the type-III seesaw scenario.

At the LHC, the typical lepton-number-violating signatures of the type-III seesaw mechanism can be $pp \to \Sigma^+\Sigma^0 \to l_\alpha^+ l_\beta^+ + Z^0 W^-(\to 4j)$ and $pp \to \Sigma^-\Sigma^0 \to l_\alpha^- l_\beta^- + Z^0 W^+(\to 4j)$ processes. A detailed analysis of such collider signatures have been done in the literature. As for the low-energy phenomenology, a consequence of this seesaw scenario is the non-unitarity of the 3×3 flavor mixing matrix N ($\approx V$) in both charged- and neutral-current interactions. Current experimental bounds on the deviation of $NN^\dagger$ from the identity matrix are at the 0.1% level, much stronger than those obtained in the type-I seesaw scheme, just because the flavor-changing processes with charged leptons are allowed at the tree level in the type-III seesaw mechanism.

8.5 Inverse and multiple seesaws

Given the naturalness and testability as two prerequisites, the double or inverse seesaw mechanism is another interesting possibility of generating tiny neutrino masses at the TeV scale. The idea of this seesaw picture is to add three heavy right-handed neutrinos $N_{\rm R}$, three SM gauge-singlet neutrinos $S_{\rm R}$ and one Higgs singlet Φ into the SM, such that the gauge-invariant lepton mass terms can be written as

$$-\mathcal{L}_{\rm lepton} = \overline{l_{\rm L}} Y_l H E_{\rm R} + \overline{l_{\rm L}} Y_\nu \tilde{H} N_{\rm R} + \overline{N_{\rm R}^{\rm c}} Y_S \Phi S_{\rm R} + \frac{1}{2} \overline{S_{\rm R}^{\rm c}} \mu S_{\rm R} + {\rm h.c.} \; , \tag{164}$$

where the μ-term is naturally small according to 't Hooft's naturalness criterion, because it violates the lepton number. After spontaneous gauge symmetry breaking, the overall neutrino mass term turns out to

be

$$-\mathcal{L}_{\rm mass} \; = \; \frac{1}{2} \, \overline{\left(\nu_{\rm L} \; N^{\rm c}_{\rm R} \; S^{\rm c}_{\rm R}\right)} \left(\begin{matrix} \mathbf{0} & M_{\rm D} & \mathbf{0} \\ M^T_{\rm D} & \mathbf{0} & M_S \\ \mathbf{0} & M^T_S & \mu \end{matrix} \right) \left(\begin{matrix} \nu^{\rm c}_{\rm L} \\ N_{\rm R} \\ S_{\rm R} \end{matrix} \right) \; , \tag{165}$$

where $M_{\rm D} = Y_\nu \langle H \rangle$ and $M_S = Y_S \langle \Phi \rangle$. A diagonalization of the symmetric 9×9 matrix $\mathcal{M}$ leads us to the effective light neutrino mass matrix

$$M_\nu \approx M_{\rm D} \frac{1}{M^T_S} \mu \frac{1}{M_S} M^T_{\rm D} \tag{166}$$

in the leading-order approximation. Hence the smallness of M_ν can be attributed to both the smallness of μ itself and the doubly-suppressed $M_{\rm D}/M_S$ term for $M_{\rm D} \ll M_S$. For example, $\mu \sim \mathcal{O}(1)$ keV and $M_{\rm D}/M_S \sim \mathcal{O}(10^{-2})$ naturally give rise to a sub-eV M_ν. One has $M_\nu = \mathbf{0}$ in the limit $\mu \to \mathbf{0}$, which reflects the restoration of the slightly-broken lepton number. The heavy sector consists of three pairs of pseudo-Dirac neutrinos whose CP-conjugated Majorana components have a tiny mass splitting characterized by the order of μ.

Going beyond the canonical (type-I) and inverse seesaw mechanisms, one may build the so-called "multiple" seesaw mechanisms to further lower the seesaw scales.

9 Non-unitary Neutrino Mixing

It is worth remarking that the charged-current interactions of light and heavy Majorana neutrinos are not completely independent in either the type-I seesaw or the type-(I+II) seesaw. The standard charged-current interactions of ν_i and N_i are already given in Eq. (34), where V is just the light neutrino mixing matrix responsible for neutrino oscillations, and R describes the strength of charged-current interactions between (e, μ, τ) and (N_1, N_2, N_3). Since V and R belong to the same unitary transformation done in Eq. (28) or Eq. (159), they must be correlated with each other and their correlation signifies an important relationship between neutrino physics and collider physics.

It can be shown that V and R share nine rotation angles (θ_{i4}, θ_{i5} and θ_{i6} for $i = 1$, 2 and 3) and nine phase angles (δ_{i4}, δ_{i5} and δ_{i6} for $i = 1$, 2 and 3). To see this point clearly, let us decompose V into $V = AV_0$, where V_0 is the standard (unitary) parametrization of the 3×3 PMNS matrix in which three CP-violating phases δ_{ij} (for $ij = 12, 13, 23$) are associated with s_{ij} (i.e., $c_{ij} \equiv \cos\theta_{ij}$ and $\hat{s}_{ij} \equiv e^{i\delta_{ij}} \sin\theta_{ij}$). Because of $VV^\dagger = AA^\dagger = \mathbf{1} - RR^\dagger$, it is obvious that $V \to V_0$ in the limit of $A \to \mathbf{1}$ (or equivalently, $R \to \mathbf{0}$). Considering the fact that the non-unitarity of V must be a small effect (at most at the percent level as constrained by current neutrino oscillation data and precision electroweak data), we expect $s_{ij} \leq \mathcal{O}(0.1)$ (for $i = 1, 2, 3$ and $j = 4, 5, 6$) to hold. Then we obtain

$$R = \begin{pmatrix} \hat{s}^*_{14} & \hat{s}^*_{15} & \hat{s}^*_{16} \\ \hat{s}^*_{24} & \hat{s}^*_{25} & \hat{s}^*_{26} \\ \hat{s}^*_{34} & \hat{s}^*_{35} & \hat{s}^*_{36} \end{pmatrix} \tag{167}$$

as an excellent approximations. A striking consequence of the non-unitarity of V is the loss of universality for the Jarlskog invariants of CP violation, $J^{ij}_{\alpha\beta} \equiv {\rm Im}(V_{\alpha i} V_{\beta j} V^*_{\alpha j} V^*_{\beta i})$, where the Greek indices run over (e, μ, τ) and the Latin indices run over $(1, 2, 3)$. For example, the extra CP-violating phases of V are possible to give rise to a significant asymmetry between $\nu_\mu \to \nu_\tau$ and $\overline{\nu}_\mu \to \overline{\nu}_\tau$ oscillations.

The probability of $\nu_\alpha \to \nu_\beta$ oscillations in vacuum, defined as $P_{\alpha\beta}$, is given by

$$P_{\alpha\beta} = \frac{\sum\limits_i |V_{\alpha i}|^2 |V_{\beta i}|^2 + 2 \sum\limits_{i<j} {\rm Re} \left(V_{\alpha i} V_{\beta j} V^*_{\alpha j} V^*_{\beta i}\right) \cos\Delta_{ij} - \sum\limits_{i<j} J^{ij}_{\alpha\beta} \sin\Delta_{ij}}{\left(VV^\dagger\right)_{\alpha\alpha} \left(VV^\dagger\right)_{\beta\beta}} \; , \tag{168}$$

where $\Delta_{ij} \equiv \Delta m^2_{ij} L/(2E)$ with $\Delta m^2_{ij} \equiv m^2_i - m^2_j$, E being the neutrino beam energy and L being the baseline length. If V is exactly unitary (i.e., $A = \mathbf{1}$ and $V = V_0$), the denominator of Eq. (168) will become unity and the conventional formula of $P_{\alpha\beta}$ will be reproduced. Note that $\nu_\mu \to \nu_\tau$ and $\overline{\nu}_\mu \to \overline{\nu}_\tau$ oscillations may serve as a good tool to probe possible signatures of non-unitary CP violation. To illustrate this point, we consider a short- or medium-baseline neutrino oscillation experiment with $|\sin\Delta_{13}| \sim |\sin\Delta_{23}| \gg |\sin\Delta_{12}|$, in which the terrestrial matter effects are expected to be insignificant or negligibly small. Then the dominant CP-conserving and CP-violating terms of $P(\nu_\mu \to \nu_\tau)$ and $P(\overline{\nu}_\mu \to \overline{\nu}_\tau)$ are

$$\begin{aligned} P(\nu_\mu \to \nu_\tau) &\approx \sin^2 2\theta_{23} \sin^2 \frac{\Delta_{23}}{2} - 2\left(J^{23}_{\mu\tau} + J^{13}_{\mu\tau}\right) \sin\Delta_{23} \; , \\ P(\overline{\nu}_\mu \to \overline{\nu}_\tau) &\approx \sin^2 2\theta_{23} \sin^2 \frac{\Delta_{23}}{2} + 2\left(J^{23}_{\mu\tau} + J^{13}_{\mu\tau}\right) \sin\Delta_{23} \; , \end{aligned} \tag{169}$$

where the good approximation $\Delta_{13} \approx \Delta_{23}$ has been used in view of the experimental fact $|\Delta m^2_{13}| \approx |\Delta m^2_{23}| \gg |\Delta m^2_{12}|$, and the sub-leading and CP-conserving "zero-distance" effect has been omitted. For simplicity, I take V_0 to be the exactly tri-bimaximal mixing pattern (i.e., $\theta_{12} = \arctan(1/\sqrt{2})$, $\theta_{13} = 0$ and $\theta_{23} = \pi/4$ as well as $\delta_{12} = \delta_{13} = \delta_{23} = 0$) and then arrive at

$$2\left(J^{23}_{\mu\tau} + J^{13}_{\mu\tau}\right) \; \approx \; \sum_{l=4}^{6} s_{2l} s_{3l} \sin\left(\delta_{2l} - \delta_{3l}\right) \; . \tag{170}$$

Given $s_{2l} \sim s_{3l} \sim \mathcal{O}(0.1)$ and $(\delta_{2l} - \delta_{3l}) \sim \mathcal{O}(1)$ (for $l = 4, 5, 6$), this non-trivial CP-violating quantity can reach the percent level. When a long-baseline neutrino oscillation experiment is concerned, however, the terrestrial matter effects must be taken into account because they might fake the genuine CP-violating signals. As for $\nu_\mu \to \nu_\tau$ and $\overline{\nu}_\mu \to \overline{\nu}_\tau$ oscillations under discussion, the dominant matter effect results from the neutral-current interactions and modifies the CP-violating quantity of Eq. (170) in the following way:

$$2\left(J^{23}_{\mu\tau} + J^{13}_{\mu\tau}\right) \Longrightarrow \sum_{l=4}^{6} s_{2l} s_{3l} \left[\sin\left(\delta_{2l} - \delta_{3l}\right) + A_{\rm NC} L \cos\left(\delta_{2l} - \delta_{3l}\right)\right] \; , \tag{171}$$

where $A_{\rm NC} = G_{\rm F} N_n/\sqrt{2}$ with N_n being the background density of neutrons, and L is the baseline length. It is easy to find $A_{\rm NC} L \sim \mathcal{O}(1)$ for $L \sim 4 \times 10^3$ km.

10 Concluding Remarks

I have briefly described some basic properties of massive neutrinos in an essentially model-independent way in these lectures, which are largely based on the book by Dr. Shun Zhou and myself [1] and on a few review articles or lectures [2]— [6]. It is difficult to cite all the relevant references. I apologize for missing other people's works due to the tight page limit of these proceedings. For the same reason I am unable to write in the cosmological matter-antimatter asymmetry and the leptogenesis mechanism, although they were discussed in my lectures. Here let me just give a few remarks on the naturalness and testability of TeV seesaw mechanisms.

Although the seesaw ideas are elegant, they have to appeal for some or many new degrees of freedom in order to interpret the observed neutrino mass hierarchy and lepton flavor mixing. According to Weinberg's *third law of progress in theoretical physics*, "you may use any degrees of freedom you like to describe a physical system, but if you use the wrong ones, you will be sorry." What could be better?

Anyway, we hope that the LHC might open a new window for us to understand the origin of neutrino masses and the dynamics of lepton number violation. A TeV seesaw might work (*naturalness*?)

and its heavy degrees of freedom might show up at the LHC (*testability*?). A bridge between collider physics and neutrino physics is highly anticipated and, if it exists, will lead to rich phenomenology.

I am indebted to the organizers of AEPSHEP 2014 for their invitation and hospitality. This work is supported in part by the National Natural Science Foundation of China under grant No. 11135009.

References

[1] Z.Z. Xing and S. Zhou, *Neutrinos in Particle Physics, Astronomy and Cosmology*, Zhejiang University Press and Springer-Verlag (2011).

[2] Z.Z. Xing, plenary talk given at ICHEP2008; Int. J. Mod. Phys. A **23**, 4255 (2008).

[3] Z.Z. Xing, Prog. Theor. Phys. Suppl. **180**, 112 (2009).

[4] H. Fritzsch and Z.Z. Xing, Prog. Part. Nucl. Phys. **45**, 1 (2000);

[5] Z.Z. Xing, Int. J. Mod. Phys. A **19**, 1 (2004).

[6] Z.Z. Xing, Lectures given at the 2010 Schladming Winter School on *Masses and Constants*, Austria, 2010; published in Nucl. Phys. Proc. Suppl. **203-204**, 82 (2010).

Proceedings of the 2014 Asia–Europe–Pacific School of High-Energy Physics, Puri, India, 4–17 November 2014, edited by
M. Mulders and R. Godbole, CERN Yellow Reports: School Proceedings, Vol. 2/2017, CERN-2017-005-SP (CERN, Geneva, 2017)

Heavy-Ion Physics

S. Gupta
Department of Theoretical Physics, Tata Institute of Fundamental Research, Mumbai, India

Abstract
Heavy-ion collisions provide the only laboratory tests of relativistic quantum field theory at finite temperature. Understanding these is a necessary step in understanding the origins of our universe. These lectures introduce the subject to experimental particle physicists, in the hope that they will be useful to others as well. The phase diagram of QCD is briefly touched upon. Kinematic variables which arise in the collisions of heavy-ions beyond those in the collisions of protons or electrons are introduced. Finally, a few of the signals studied in heavy-ion collisions, and the kind of physics questions which they open up are discussed.

Keywords
Lectures; quantum chromodynamics; heavy-ion collisions; quark-gluon plasma; chiral symmetry; jet quenching.

1 Why study heavy-ion collisions

The universe started hot and small, and cooled as it expanded. Today vast parts of the universe are free of particles, except for photons with energy of about 3 Kelvin or lower. This energy scale is so far below the mass scale of any other particle that no scattering processes occur in this heat bath of photons. So we may consider them to be free.

This was not always so. Earlier in the history of the universe the temperature T was comparable to, or larger than, the mass scales of many particles. As a result particle production and transmutations were common. In those circumstances would it be correct and useful to treat this fluid as an ideal gas? Such a gas cannot give rise to freeze out, phase transitions or rapid crossovers, and transport. We see the signatures of several such phenomena today, so we know that the ideal gas treatment would not work at all times.

In the early universe many of the component particles of the fluid were relativistic. Since we wish to describe particle production processes in this fluid, we are forced to use quantum field theory at finite temperature to describe the contents of the early universe. The main theoretical tools required to study thermal quantum field theory (TQFT) are effective field theory (which includes hydrodynamics and transport theory) and lattice field theory. Perturbation theory plays a limited but very important role, due to our detailed understanding of the technique. In order to test the formulation of TQFT, we need to think of experiments which can be performed easily.

Experimental tests of TQFT in the electro-weak sector turn out to be unfeasible. Initial states made of leptons may achieve energy densities of the order of $1/\text{fm}^4$. However, mean-free paths due to electro-weak interactions are of the order 100 fm. So it is very hard to thermalize this energy density. Initial states of hadrons, on the other hand, have mean-free paths of the order of 1 fm, so the initial energy may be converted into thermal energy. By using heavy-ions, one can increase the initial volume significantly, and so improve the chances of producing thermalized matter. This is why heavy-ion collisions (HICs) are used to test TQFT.

The objects of experimental study should be as many as possible, in order to subject TQFT to as many tests as can be conceived. The most important phenomena are transport properties: the electrical conductivity (important for the freezeout of photons), viscosity (responsible for entropy production), the

https://doi.org/10.23730/CYRSP-2017-002.219

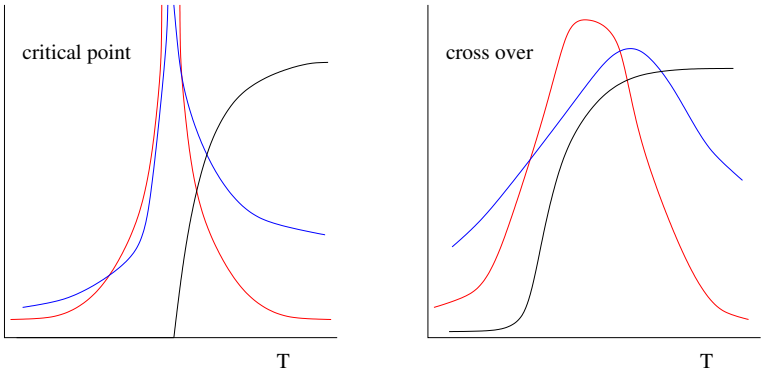


Fig. 1: At a critical point order parameters change abruptly, and specific heats and susceptibilities may have singularities. The location of the singularity is unique. At a crossover there are no singularities. The order parameters may have a large continuous change. It is possible that specific heats and susceptibilities peak as the temperature changes. The locations of maximum slope of the order parameter, or the peaks of susceptibilities, generally depend on the observable chosen.

speed of sound, the equation of state and so on. But perhaps the most interesting objects of experimental study are the possible phase transitions and crossovers associated with the symmetries of the standard model. Corresponding to every global symmetry there is a chemical potential. So the phase diagram of the standard model has high dimensionality and potentially many phases. Experiments which are feasible in colliders can reach only a small fraction of the phases.

2 Symmetries and states of QCD

Phase diagrams display the conditions under which global symmetries are broken or restored. Heavy-ion collisions explore the phase diagram of QCD. The global symmetries of this theory are chiral $\mathrm{SU}_L(N_f)$ $\times$ $\mathrm{SU}_R(N_f)$ $\times$ $\mathrm{U}_B(1)$, where N_f is the number of flavours of light quarks, the subscripts L and R stand for left and right chirality, and B for baryon number.

Chiral symmetry is explicitly broken by quark masses. QCD contains a scale, Λ_{QCD}. Quarks with masses larger than Λ_{QCD} are far from the chiral limit. The strange quark mass, m_s, is near the scale of Λ_{QCD}, and it is a detailed question whether treating it as nearly chiral helps in understanding the phenomenology of strong interactions. The up and down quark masses are much lighter than Λ_{QCD} and it is useful to treat them as nearly chiral.

The resulting $\mathrm{SU}_L(2)$ $\times$ $\mathrm{SU}_R(2)$ *chiral symmetry is spontaneously broken* down to SU(2) isospin symmetry in the vacuum. Signals of this symmetry breaking are the fact that the QCD vacuum contains a non-vanishing chiral condensate, $\langle\overline{\psi}\psi\rangle$, and that pions are massless. Departures from chirality are important and treated in chiral perturbation theory [1]: the most important result is that pions get a mass proportional to the square root of the quark mass.

As the temperature of the vacuum is raised, keeping the baryon number and charge densities at zero, the condensate changes to a very small value, proportional to the quark mass. From thermodynamic arguments, models, and lattice QCD computations it is known that the change is gradual (see Figure 1). One may try to characterize a temperature where this crossover happens, but it is a conventional number [2]. The *crossover temperature*, T_c, depends upon which physical quantity is examined, but it is perfectly well-defined after a choice is made[1] We make the choice that T_c is given by the peak of the Polyakov loop susceptibility.

[1]Another example of a crossover is the formation of a glass by cooling of liquid silica. The glass transition temperature depends on what measurement one makes on the sample of the glass.

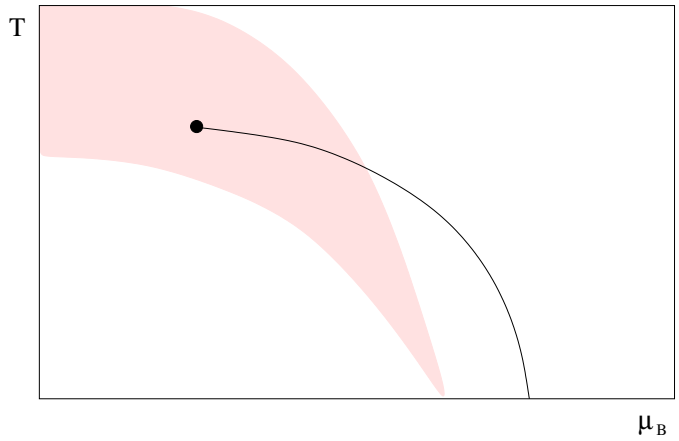


Fig. 2: The phase diagram of QCD in the T-μ_B plane as seen with the chiral condensate [3]. A line of first order phase transitions (black line) ends in the QCD critical point (black dot). The fireball produced in a heavy-ion collision lives in a track within the shaded domain. The lower edge of this domain is called the freezeout curve. The domain is traced out by tracks of the history of the fireball as the collider energy changes. For small energies, the domain ends at $T = 0$ and a chemical potential corresponding to nuclear matter. As the energy increases, the domain moves to small μ_B. The logic of the beam energy scan is that this domain is likely to include the QCD critical point.

The quark number for each of the N_f flavours is conserved. For the study of the phase diagram we need to keep in mind the up and down quark numbers (or, equivalently, the baryon number and the net isospin). A grand-canonical ensemble for QCD would then need two chemical potentials, μ_u and μ_d (or μ_B and μ_I), and the temperature T: so the phase diagram is three dimensional. As a first approximation one treats the up and down quark masses to be equal, and examines the phase diagram in the two dimensional slice with T and μ_B [3], and independently, of that in T and μ_I [4]. There has been little study of the more complete (and complicated) phase diagram [5].

The phase diagram in T and μ_B for small μ_B and non-zero light quark mass (see Figure 2) was first investigated in [3]. At small μ_B the states of QCD are distinguished by the value of the chiral condensate. In the low T state it is large, but becomes small at $T > T_c$. At sufficiently high T the dependence of the condensate on μ_B is computable, and shows a gradual variation. However, various arguments lead to the expectation that at low T as one changes μ_B there is a first order phase transition signalled by an abrupt change in the condensate. The thermodynamic Gibbs' phase rule [6] then tells us that the phase diagram has a line of first order transition. As we already discussed, this cannot hit the $\mu_B = 0$ axis or rise to $T \to \infty$. So it must end somewhere. The end point is a second order phase transition, called the QCD critical point.

The actual location of the curve of first order phase transition curve and the critical end point can only be predicted by a non-perturbative computation, *i.e.*, through *lattice QCD simulations* [7]. However, a computation at finite μ_B requires an extension of known techniques because of a technical problem known as the *fermion sign problem*. Many such methods have been proposed, and many are being explored [7]. It would be fair to say that developing such techniques is one of the most active areas in lattice gauge theory today.

Till now extensive computations in QCD with varying lattice cutoffs, spatial volumes and quark masses has been possible using only one particular method, involving the Taylor expansion of the pressure in powers of μ_B. As a result the information available until now is fairly limited, and one would hope that the future brings alternative computational schemes. The current best estimate of the location

of the critical point is [8]

$$T^E \simeq (0.94 \pm 0.01) T_c \qquad \text{and} \qquad \mu_B^E \simeq (1.68 \pm 0.06) T^E. \tag{1}$$

Methods have also been developed to compute the equation of state, the bulk compressibility, and the speed of sound in several parts of the phase diagram [8].

Two more aspects of the phase diagram of QCD are interesting, but cannot be described here. One is the temperature dependence of the axial anomaly. This has been keenly investigated in recent years [9]. The other is the phase diagram of QCD in a strong and constant external magnetic field. This has also generated much work recently [10].

3 General conditions in heavy-ion collisions

I turn now to heavy-ion collisions, which is the experimental system that can test the computations we discussed briefly in the previous section. In this section I touch upon three related questions: whether thermal matter is produced, what its flavour content is likely to be, and how one can control the energy content of this matter.

The object of study in heavy-ion collisions is the (hopefully) thermalized matter in the final state. In high energy colliders matter is always formed. In sufficiently hard pp collisions, for example at the LHC, even soft physics contains enough energy to create W/Z bosons, not to speak of hadrons. The mere production of large amounts of hadronic matter is not of interest. What we need to know is whether this matter re-interacts with sufficient strength to thermalize. In the language of particle physics this is about final state effects.

In order to understand the time scales involved it is sufficient to run through a simple kinetic theory argument. Let the two-body scattering cross section be σ. Taking the number density of particles in the final state to be n, one can write the mean free path as

$$\lambda \propto (n\sigma)^{-1}, \tag{2}$$

If the dimensionless number $1/(\lambda \sqrt[3]{n}) = \sigma/n^{2/3} = \mathcal{O}(1)$ then the mean free path is of the same order as the mean separation between particles. In this case, final state collisions are numerous, and matter may come into *local thermal equilibrium*.

When $\sqrt{S} \simeq 20$ GeV, we know that jets are rare. As a result, we can take the final state particles to be hadrons, so that $\sigma \simeq 40$ mb. In this case $n \geq 5/\text{fm}^3$ may be sufficient for the final state to thermalize. This number density cannot be reached in collisions of protons. However, heavy-ion collisions increases n by some power of A, so heavy-ion collisions at this energy may thermalize. At the LHC, n is large, so thermalization is easier. Even high multiplicity pp collisions may then thermalize. The thermalized system arising from these collisions is the *fireball* which is the object of study in heavy-ion collisions.

This treatment is sufficient for building intuition, but a quantitative analysis of thermalization is more complex. The rapid expansion of the fireball implies that simple kinetic theory does not suffice, and the theoretical framework becomes more complex. Some relevant references are collected here [11].

The flavour content of the fireball is needed in many analyses. Again, simple arguments are sufficient to gain a quick intuition about this. The flavour quantum numbers of the incoming hadrons are essentially contained in hard (valence) quarks. At large $\sqrt{S}$, the asymptotic freedom of QCD guarantees that our intuition about Rutherford scattering holds, and these valence quarks do not undergo large angle scattering. As a result, the incoming quantum numbers are mostly carried forward into the fragmentation region. In terms of the pseudo-rapidity

$$\eta = \frac{1}{2} \log \tan \theta, \tag{3}$$

(where θ is the scattering angle) the *fragmentation region* is the region of large $|\eta|$, and is called so because (classically) one finds the unscattered fragments of the initial particles here.

Although the valence partons individually contain large momenta, there are only three of them in a baryon. Soft (sea) partons are much more numerous. As a result, quite a significant fraction of the energy is carried by all the soft partons together. These generally scatter by large angles and so stay in the *central rapidity region* (*i.e.*, the region with $|\eta| \simeq 1$). If this matter approximately thermalizes, then it makes the fireball which is the object of heavy-ion studies. The net-baryon and flavour content is small, the energy content increases with $\sqrt{S}$. At high energies the central and fragmentation regions are expected to be well separated, *i.e.*, one expects few hadrons in the intermediate region between them.

At $\sqrt{S} \simeq$1–10 GeV, baryon interactions cannot be analyzed in terms of quarks. In this regime the fireball may contain baryon and other flavour quantum numbers. The distinction between fireball (central) and fragmentation region may be weak.

In the collision of point-like particles in quantum theory, the observables are the number of particles (or energy-momentum) hitting the detector at any angle θ. The only control parameter is the center of mass energy, $\sqrt{S}$. In collisions of extended objects, there is another control parameter: the impact parameter, b. This measures the separation between the centers (geometrical centers, centers of energy) of the colliding objects. However, b cannot directly be measured in an experiment.

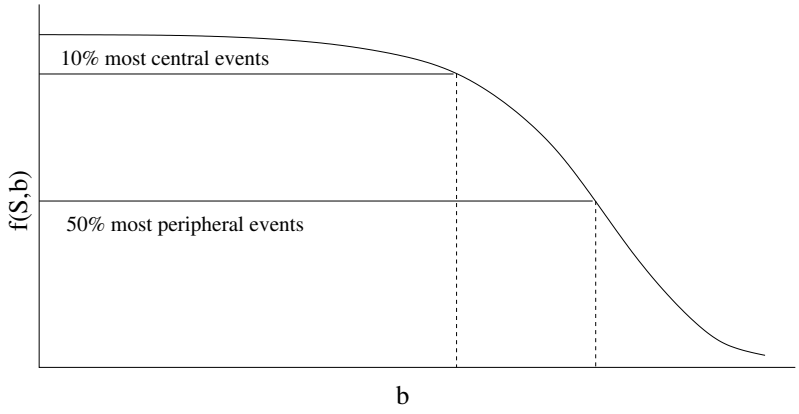


Fig. 3: The fraction of the total cross section can be used to define centrality classes. Different models of nuclear densities will map a centrality class to different impact parameter ranges.

Instead we perform the following analysis. The total nucleus-nucleus cross section depends only on the energy, so one has

$$\sigma(S) = \int_0^\infty db \frac{d\sigma(S)}{db}. \tag{4}$$

Since cross sections are non-negative, the fractional cross section

$$f(S,b) = \frac{1}{\sigma(S)} \int_b^\infty dB \frac{d\sigma(S)}{dB},$$

decreases monotonically as b increases from zero to infinity (see Figure 3). As a result, an experimentally determined histogram of f would determine b uniquely, provided one knows the functional form of $f(\sqrt{S}, B)$. This is not yet computable from QCD, so one has to make models.

The simplest, and oldest, model is called the *Glauber model*. In this, one assumes that the nucleus-nucleus collision is described by independent nucleon-nucleon collisions. The nucleons are distributed in each nucleus according to the density determined by low-energy electron nucleus collisions. Models which incorporate more phenomenology have also been developed; see [12] for more information. It has been realized in recent years that the lumpy distribution of nucleons in the initial state (see Figure 4) cannot always be averaged over, but must be taken into account in these models.

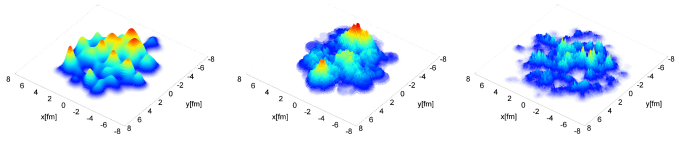


Fig. 4: Transverse energy profiles, 0.2 fm after the collision, in three models of initial states [13]. The coarse-grained average of these distributions should give the nuclear density known through low-energy experiments. These relativistic experiments capture the quantum fluctuations in the initial nuclear wave-function.

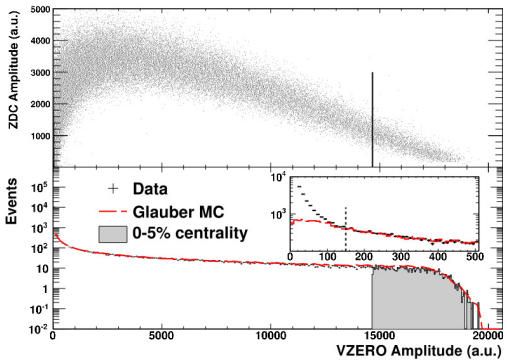


Fig. 5: One of the definitions of centrality used by the ALICE collaboration [14] uses the histogram of measurements in the VZERO module. The experiment determines its correlation with the energy deposition in the ZDC.

As one sees, the connection between the impact parameter and the percent of total cross section is indirect and model dependent. Also, when one realizes that the positions of nucleons inside the nuclei fluctuate from one event to another, it is clear that the notion of an impact parameter, and even the size and shape of a nucleus, are merely averaged quantities. For experimental purposes what is necessary is to classify events according to the degree of centrality. For this it is enough to define centrality by any measure which changes monotonically with b; for example, by multiplicity, zero degree calorimetry, *etc.*. Care is needed to relate these measures to each other through careful analysis of the data. One such analysis is shown in Figure 5.

4 Hard probes

One thinks of the LHC as an arena of hard QCD, *i.e.*, of processes which convert the partons contained in protons into jets, heavy quarks, W/Z bosons, hard γ, H and so on. The typical momentum scale in these processes is of the order $Q \simeq \langle x \rangle \sqrt{S} \simeq 500$ GeV. Final state interactions are suppressed in pp collisions because of two reasons. Firstly, the dense hadronic debris are separated from probes by large angles, $\Delta\eta$. Secondly, the energy scale of any remaining hadronic activity in the central rapidity region is small: $\langle E_T \rangle \simeq \Lambda_{QCD} \simeq 0.3$ GeV.

In heavy-ion collisions, the first argument can still be supported. However, the second argument may fail if the number density of particles, n, is large enough. Let us make an estimate by assuming, as before, that $n = 5/\mathrm{fm}^3$. We know that the actual value of n at the LHC is larger, so our argument will be overly conservative. Assume that the jet cone has radius[2] $R = 0.2$, and that it travels about $\ell = 10$

[2]The radius of a jet cone is defined to be $R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ where we take $\Delta\eta$ and $\Delta\phi$ to be the jet opening angles.

fm through the fireball of soft particles. Then the net energy in the soft hadrons it can interact with is

$$\mathcal{E} \simeq \langle E_T \rangle n R \ell^3 \simeq 300 \text{ GeV}, \tag{5}$$

where we have made a conservative estimate that the average transverse energy of the particles is $\langle E_T \rangle \simeq 0.3$ GeV. Since this is comparable with the initial energy, final state interactions become important. An interesting consequence which we discuss here is jet quenching.

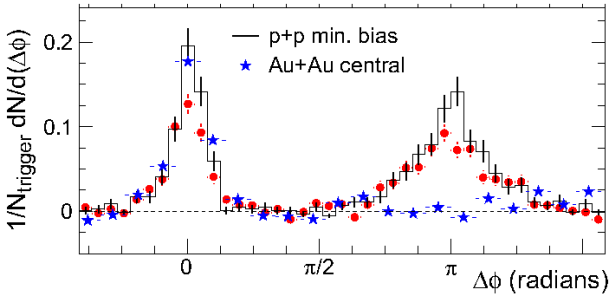


Fig. 6: Comparison of two-particle azimuthal distributions for central d+Au collisions (circles) to those seen in p+p (histogram) and Au+Au collisions (stars). The respective pedestals have been subtracted.

When a jet evolves through a medium, interactions and radiation would tend to deplete its energy [15]. This simple idea is called *jet quenching*. The basic fact of jet quenching was beautifully demonstrated by the STAR collaboration in BNL [16] in the plot given above. At $\sqrt{S} = 200$ GeV jets are not very well developed, and one must use high-p_T hadrons as proxies. STAR triggered on events where there is a high-p_T hadron, and looked at the angular distribution of the next highest-p_T hadron. In p+p collisions they found a peak 180 degrees away (see Figure 6). If the trigger hadron can be assumed to come from a jet, then the backward peak comes from an away-side jet which balances the momentum. This was also seen in d+Au collisions, thus demonstrating that initial state parton effects in heavy nuclei do not wash away this peak. In Au+Au collisions they found no peak in the backward direction: implying that the away-side jet is hugely quenched[3].

A measure of the quenching is provided by a comparison of the number of jets of a given momentum in heavy-ion and proton collisions

$$R_{AA}(b, y, p_T) = \frac{1}{T_{AA}(b)} \frac{d^3 N_{AA}}{db dy dp_T} \left(\frac{d^2 N_{pp}}{dy dp_T} \right)^{-1}. \tag{6}$$

Here T_{AA} is an estimate of the number of proton pairs interacting in AA collisions, and is usually extracted from a model, *e.g.*, the Glauber model. The numerator depends on collision centrality whereas the denominator does not. Energy is tremendously more likely to flow from the jet into the low-momentum particles in the medium (computations reveal this in phase space factors). As a result, one would generally expect R_{AA} to be less than unity.

Since a basic input into jet-quenching is T_{AA}, it is important to constrain this through experiment. The production of high-p_T photons or W/Z bosons provides this calibration. Since the vector bosons have no strong interactions, the comparison of semi-inclusive single boson production cross sections in pp and AA cross sections can directly measure T_{AA}. One of the first attempts [17] to constrain this is shown in Figure 7. Small isospin corrections, shadowing, and initial re-scattering effects must also be taken into account more accurately in order to improve these constraints.

[3] Since the near-side jet is used as a trigger, the event sample is of those in which this is not completely quenched.

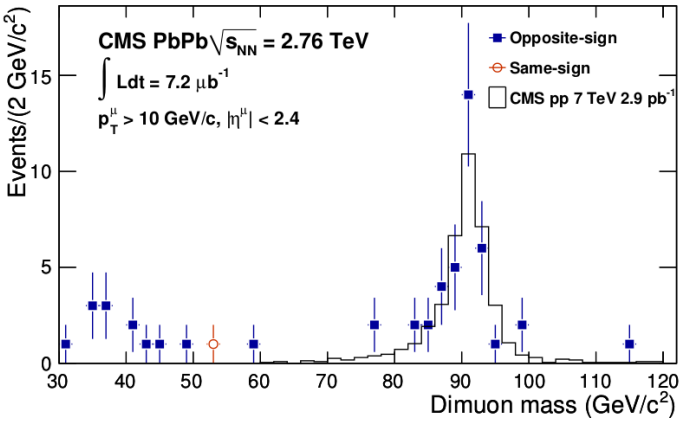


Fig. 7: The first attempt to constrain T_{AA} from experiment.

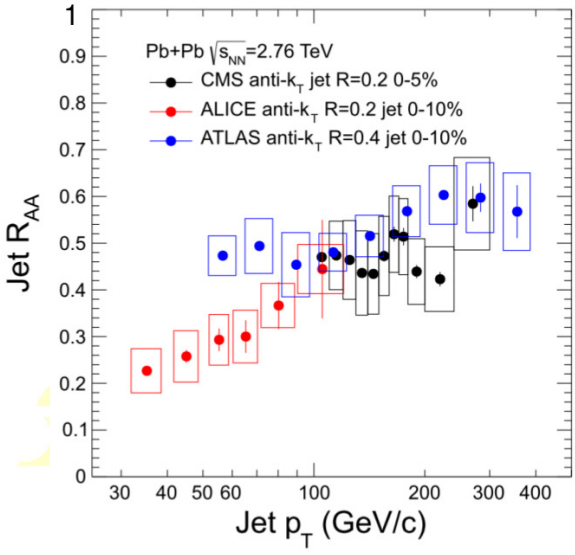


Fig. 8: A recent compilation of data on R_{AA} [18].

From the observations of R_{AA} (see Figure 8) one can extract a measure of the p_T change of the jet per unit distance travelled within the plasma:

$$\hat{q} = \frac{1}{L} \int \frac{d^2 p_T}{(2\pi)^2} \, p_T^2 \, P(p_T, L).$$

Most attempts to extract this from data give $\hat{q} \simeq$ 1–2 GeV^2/fm, *i.e.*, in the range of interest, $\hat{q}/T^3 \simeq$4–5. One should be able to extract this number for QCD, but it turns out to be a vexing problem. There are two main methods to handle problems in QCD: perturbative QCD is used to compute processes where all momenta involved are large, and lattice QCD provides a tractable computational method when all momenta are small. Jet quenching couples a large momentum object (the jet) to low-momentum objects (the medium). Nevertheless there have been attempts to compute this in QCD using weak-coupling expansions [19] or, more recently, lattice QCD [20]. There are also computations in cousins of QCD which have $\mathcal{N} = 4$ supersymmetry in the limit of large N_c [21].

R_{AA} is just the simplest of experimental variables which can be constructed. In order to understand how the medium steals energy and momentum from the jet one should also understand medium

modification of rapidity and angular correlation, momentum imbalance between reconstructed jets, fragmentation functions, and jet substructure.

5 Flow

Once the fireball reaches local thermal equilibrium a much slower process begins of transport of energy, momentum, and other conserved quantities through the fireball. This is the hydrodynamic regime [22]. Tests of hydrodynamics involve the study of quantities which are called flow coefficients [23]. In order to understand what these are, we need to think again about the geometry and kinematics of the collisions.

In the collision of point-like particles, there is a rotational symmetry around the beam axis. As a result, cross sections or particle production rates depend only on the scattering angle θ (or equivalently, on η or the rapidity y) and the transverse momentum, p_T. Kinematically, there is only one initial vector in the center of mass (CM) frame of the problem, the initial momentum $\mathbf{k}$ of one of the particles (the other particle has momentum $-\mathbf{k}$). Final state momenta see only the angle from $\mathbf{k}$, which is θ, and the transverse projection p_T.

In heavy-ion collisions, there is a second initial vector: $\mathbf{b}$, which is the line between the centers of the nuclei. The existence of such a vector, not collinear with $\mathbf{k}$, means that the azimuthal symmetry around $\mathbf{k}$ is broken in the initial state, and final state momentum distributions may depend on angles the final momentum makes with both $\mathbf{k}$ and $\mathbf{b}$ as well as p_T. Conventionally, these distributions are given in terms of η, p_T, and the azimuthal angle ϕ. The two vectors $\mathbf{k}$ and $\mathbf{b}$ lie on a plane which is called the *reaction plane*. This breaking of cylindrical symmetry also occurs in proton-nucleus collisions, since the proton can meet the nucleus with a non-vanishing impact parameter. In very high energy collisions, the increasing proton-proton cross section implies that the swollen protons can also be treated similarly. One may already be seeing such effects in the sample of extreme high multiplicity events in pp collisions at the LHC.

The *flow coefficients* are the Fourier transforms of velocity distributions with respect to ϕ [24]. The n-th Fourier coefficient is denoted by the symbol v_n. These are normally taken at $y = 0$ not only because of the limited rapidity coverage of heavy-ion detectors, but also because one expects the fireball to be well-separated from the fragmentation region. Nevertheless, studying the rapidity dependence of flow coefficients is of some interest. The study of the k_T dependence of the v_n is of great interest.

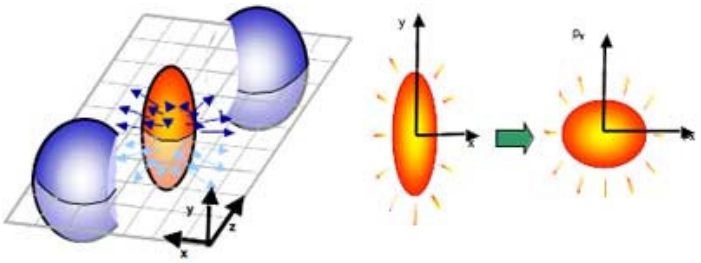


Fig. 9: The geometry of elliptic flow.

Clearly, the reaction plane in different collisions can rotate around the beam axis, so single particle distributions will recover azimuthal symmetry when averaged over events. Although the overall orientation of the reaction plane is forgotten on the average, the relative angles between two particles remembers the difference from the reaction plane. So, in order to see the flow coefficients one has to construct the angular correlations of two or more particles.

In the collision of symmetric nuclei, $\mathbf{b}$ and $-\mathbf{b}$ seem to be completely equivalent. As a result the two sides of the reaction plane seem to be completely symmetric. This implies that only the even flow coefficients, v_2, v_4, *etc.*, are non-vanishing (*elliptic flow* is the name given to v_2). However, when one

studies the flow event-by-event (E/E) one has to take into account the fact that the positions of nucleons inside the nuclei may fluctuate. Then there may be more nucleons on one side of the plane, so breaking this orientational symmetry around the reaction plane, as a result of which odd harmonics may exist. Currently there are studies of the *directed flow* v_1, *triangular flow* v_3, and even the coefficient v_5. The flow coefficients yield a combination of information on the initial state and the evolution of fireball. E/E fluctuations of flow coefficients yield more refined information on the initial state [25]

It is claimed that the observations of v_2 imply the formation of locally thermalized matter in heavy-ion collisions. Although this argument is technical it is easy to understand this intuitively. In the off-center collisions of nuclei the colliding region is a pellet. Particles formed in the initial collisions have distributions which have positional anisotropy, ϵ_n, the pellet being long in one direction (see Figure 9). The generation of v_n involves transforming ϵ_n into momentum anisotropy. This is impossible unless there is hadronic re-scattering. Also, the measurements of v_2 show that the momentum is larger in the direction in which the original position distribution was squeezed. This is hydrodynamic flow, since that is driven by pressure gradients, and the gradients in the shorter direction are larger.

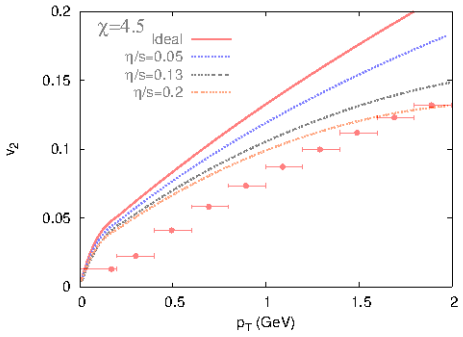


Fig. 10: Predictions for elliptic flow from ideal hydrodynamics compared to data. See the text for a discussion.

The technical question here is how well does hydrodynamics explain the observed v_2. Ideal hydrodynamics, *i.e.*, hydrodynamics without dissipation, is a toy model which is often used to understand general features of data. This already come within a factor two of the data, and shows the same p_T dependence as the observations (see Figure 10). It also fails in the "right" direction, in that it over-estimates v_2. Dissipation would clearly reduce the predictions, and bring it closer to observations [26]. One of the first results (Dusling and Teaney in [26]) is shown in the figure. Intense work continues to be done in understanding the implications of the data and the constraints from QCD [27].

6 Chemical composition of the final state

The most easily observed quantities have to do with the final state. The basic observables are the spectra of identified particles. The multiplicity of each type of particle, $\pi^{\pm}$, $K^{\pm}$, *etc.*, is called its yield. This is the integral over the spectrum. Relative yields of hadrons is the outcome of *hadron chemistry*, *i.e.*, inelastic re-scattering in the final state. Examples are,

$$p + \pi^- \leftrightarrow n + \pi^0, \qquad p + \pi^- \leftrightarrow \Lambda + K^0, \tag{7}$$

The rates of such predictions determine whether hadron chemistry comes to chemical equilibrium.

As the fireball evolves, eventually mean-free paths or relaxation times become comparable to the size or expansion rate. When this happens, local thermal equilibrium can no longer be maintained, and hydrodynamics cannot be supported. Then the components of the fireball are said to *freeze out*. In principle freeze out could occur either before the fireball cools into hadrons or after. Under normal circumstances, *i.e.*, if the thermal history of the fireball does not take it near a phase transition, then it seems

to freeze out in the hadronic phase. This implies that the later stages of hydrodynamics require the equation of state of a hadronic fluid. Since hadrons are massive, inelastic collisions, *i.e.*those which change the particle content of the fireball, require larger energies than elastic collisions. As a result, *chemical freezeout*, *i.e.*, fixing of the hadron content of the fireball, may occur earlier than *kinetic freezeout*, *i.e.*, fixing of the phase space distribution of hadrons.

In the fireball the particles interact strongly enough that a temperature is maintained. However, at freeze out the interactions stop abruptly. So all hadrons emitted by the fireball at freeze out can be assumed to be an ideal gas of particles coming from a source whose temperature is set at the freeze out. This simple approximation, which goes by the name of the *hadron resonance gas model* has had remarkable phenomenological success [28]. However, recent measurements at the LHC (ALICE collaboration [28]) and a more careful look at RHIC results shows interesting discrepancies which imply that this model needs to be improved.

At early times, the fireball is a reactive fluid whose description requires coupling of hydrodynamics with diffusion and flavour chemistry. The reaction rates depend on local densities as well as rates of mixing due to fluid movement, known as advection, as well as diffusion. In order to make quantitative predictions, one must first understand whether advection or diffusion is more important in bringing reactants together. This is controlled by *Peclet's number*

$$\mathrm{Pe} = \frac{Lv}{D} = \frac{Lv}{\xi c_s} = \mathrm{Kn}\, M, \tag{8}$$

where L is a typical macroscopic distance within the fireball over which we wish to compare advection and diffusion, v a typical flow velocity, ξ is a typical density-density correlation length and c_s is the speed of sound. The diffusion constant, $D \simeq \xi c_s$. We have also used the notation for the Mach number of the flow, $M = v/c_s$, and the Knudsen number, $\mathrm{Kn} = L/\xi$. When $\mathrm{Pe} \ll 1$ diffusion is more rapid than advection; when $\mathrm{Pe} \gg 1$ advection is more rapid [29].

Peclet's number defines a new length scale in the fireball, this is the scale at which advection and diffusion become comparable—

$$L \simeq \frac{\xi}{M}. \tag{9}$$

Since longitudinal flow has $M \leq \sqrt{3}$, then taking ξ to be approximately the Compton wavelength of a particle, we find that for baryons, $L \simeq 0.3$ fm and for strange particles, $L \simeq 0.5$ fm. This implies that advection may be important in chemical processes occurring in the early stages of the evolution of the fireball, but over most of its history, the availability of reactants is governed by diffusion.

Once the reactants have been brought together we can ask whether one or the other reaction channel is available. If the reactions are slower than the time scale of transport, then we may consider the fireball to be constantly stirred. It is then enough to examine chemical rate equations. In this approximation, a toy model which takes into account only pion and nucleon reactions is:

$$\begin{aligned}
\dot{p} &= -\gamma(p\pi^0 - n\pi^+) - \gamma'(p\pi^- - n\pi^0) + \cdots, \\
\dot{n} &= \gamma(p\pi^0 - n\pi^+) + \gamma'(p\pi^- - n\pi^0) + \cdots, \\
\dot{\pi}^0 &= -\gamma(p\pi^0 - n\pi^+) + \gamma'(p\pi^- - n\pi^0) + \cdots, \\
\dot{\pi}^+ &= \gamma(p\pi^0 - n\pi^+) + \cdots, \\
\dot{\pi}^- &= -\gamma'(p\pi^- - n\pi^0) + \cdots.
\end{aligned}$$

Here the label for a particle denotes the density of that particle. The rate constants γ and γ' can be deduced from experimental measurements of cross sections. The equilibrium concentrations are given by

$$\frac{p}{n} = \frac{\pi^+}{\pi^0} = \frac{\pi^0}{\pi^-} \quad (= \zeta), \tag{10}$$

where ζ is the isospin fugacity. Since $\pi^+/\pi^- = \zeta^2$, if we set $\zeta \simeq 1$, then $\mu_I = T \log \zeta \simeq 0$. Even in this simple limit of a very rapidly stirred fireball, a more realistic model contains all possible reactions between many species of particles, of which many cross sections are unmeasured. As a result, a detailed model is out of reach and one must develop simplified models which catch as much of the physics as the state of the data justifies.

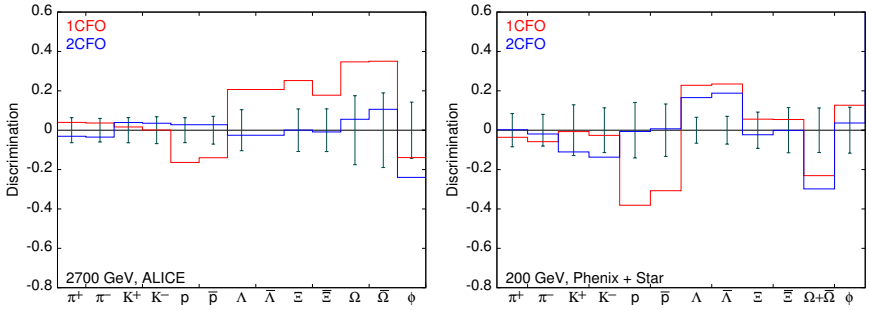


Fig. 11: Comparison of model predictions with data in terms of the discriminant (model-data). The closer this is to zero, the better the model. The error bars show the error on the data and set the scale of what is acceptable mismatch between data and model. At all energies, 2CFO works better than 1CFO.

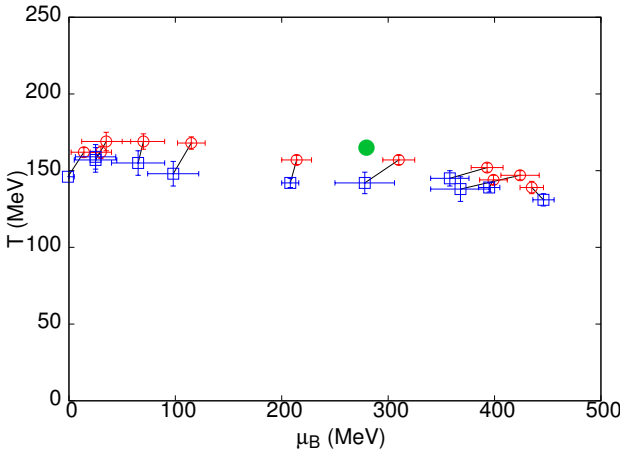


Fig. 12: The freeze out points obtained in 2CFO (from Chatterjee *et al.*, [30]). The strange freezeout point is shown with circles, and the non-strange with squares. The two freeze out points at the same $\sqrt{S}$ are joined with a line. The large filled circle is the estimated location of the QCD critical point from lattice computations [8].

In order to set up such a model, we consider flavour changing reactions. Strangeness changing processes seem to naturally split into subgroups. *Indirect transmutations* of K and π involve strange baryons in reactions such as $\Omega^- + K^+ \leftrightarrow \Xi^0 + \pi^0$. These have very high activation thresholds. *Direct transmutations* can proceed through the strong interactions such as $K^+ + K^- \leftrightarrow \pi^+ + \pi^-$. These are OZI violating reactions; slower than generic strong-interaction cross sections. Direct transmutations through weak interactions are not of relevance in the context of heavy-ion collisions. As a result, there is no physics forcing K and π to freezeout together. However K and ϕ are resonantly coupled, so they may freeze out together [30]. On the other hand, isospin changing processes (the model in eq. 10) require extremely low activation temperatures, and may persist till later.

One can capture this information into a HRG model with two freezeout points: one for the strange hadrons and ϕ (since this is resonantly coupled to the $K^\pm$ channel), another for non-strange hadrons. We

could call this model 2CFO [30] in contrast with the usual HRG model with a single freeze out (1CFO). A comparison of measurements and best fit model predictions is shown in Figure 11.

Interestingly, the introduction of two freeze out points allows one to do away with some unphysical features of the freeze out model 1CFO. In most such models there is a mismatch between strange and non-strange baryon production, which is fixed by having a fugacity factor which changes the occupancy of strange hadrons. This factor cannot be justified within an ideal gas picture, nor does it vary smoothly or monotonically as $\sqrt{S}$ is changed. Such nuisance parameters no longer appear within the 2CFO scheme.

The success of the 2CFO scheme implies that as one introduces more of the hadron dynamics into the freezeout process, the ability to describe the data improves. This justifies our belief that a proper description of reactive transport should be able to give a good description of the final observed yields.

The freeze out temperatures and chemical potentials in 2CFO are shown in Figure 12. Also shown there is the position of the critical point of QCD determined in lattice studies [8]. The freeze out curves pass close to the QCD critical point, making it plausible that a study of the final state as one scans in $\sqrt{S}$ can reveal signals of this very interesting prediction of QCD. This is the rationale for the RHIC Beam Energy Scan (BES) program and for planned future experiments in GSI and JINR.

Experiments also measure the yields of heavy-quarkonia. In particular, the yield of the Υ family of mesons in AA collisions at LHC differs significantly from that in pp collisions at same $\sqrt{S}$. This is usually reported in terms of an R_{AA} for the meson. Since the quark mass is large, $M \gg T \simeq \Lambda_{QCD}$, one may expect that the production of quarkonia is a hard process. However the binding energy is of the order of the temperature, $B \simeq T$, so we may expect large thermal effects as the cause of the change between AA and pp collisions [32].

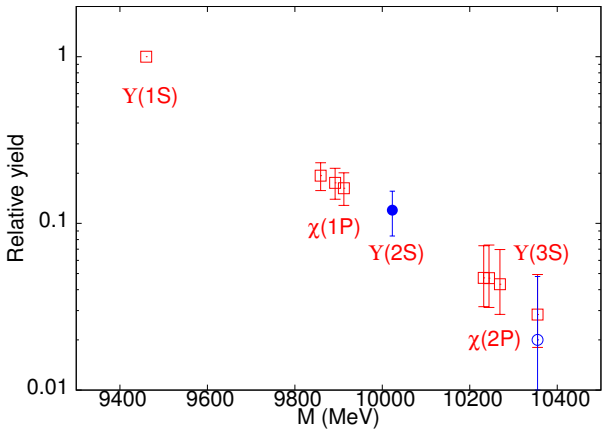


Fig. 13: Measured (filled circles) and predicted (unfilled squares) suppression in the bottomonium family using a simple thermal model for freeze out [36].

Thermal lattice QCD computations show that the highest mass resonances, which are the least bound, are more easily disrupted at any given temperature [33]. This observation led to the formulation of a key observation called *stepwise suppression*, *i.e.*, as $\sqrt{S}$ is increased R_{AA} of the higher resonances drop below unity, roughly in the order of the binding energy [34]. If this works, then at a sufficiently high temperature it should be possible to use a thermal model to understand the relative yields of the Υ family of mesons using the variables

$$r[\Upsilon(n\ell)] = \frac{dN_{AA}^{\Upsilon(n\ell)}}{dydp_T} \left(\frac{dN_{AA}^{\Upsilon(1S)}}{dydp_T} \right)^{-1} . \tag{11}$$

The thermal model involves only a single parameter: the freeze out temperature of this family of mesons.

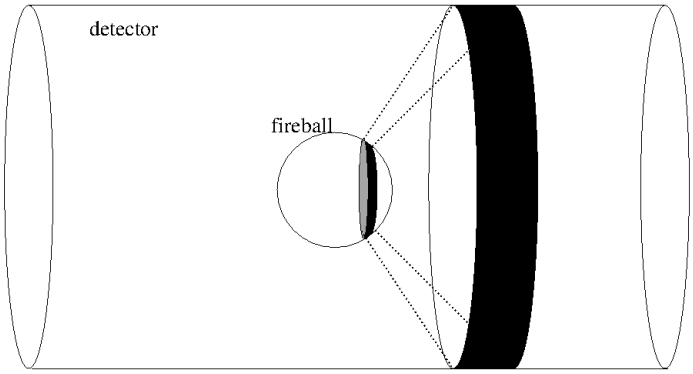


Fig. 14: Selecting a volume in the fireball by selecting detector cuts works best when the momentum of particles in the volume are not totally arbitrary. Since hydrodynamics works, we know that in a small volume these momenta are aligned with the local fluid momentum, being smeared only by an amount of order T.

The first data from the LHC [35] is fitted [36] well by

$$T_f^{\Upsilon} = 222^{+28}_{-29}\,\mathrm{MeV}. \tag{12}$$

It will be interesting in future to see whether other members of the bottomonium family confirm this picture. Future data on $r[\Psi(n\ell)]$ will also provide useful tests. More detailed dynamical models [37] predict many more details of the kinematics of quarkonium suppression.

7 Fluctuations

Is the ensemble of heavy-ion collision events captured by a detector related to the ensembles required to study the thermodynamics of strong interactions?

The least restrictive ensemble for the study of bulk matter is the microcanonical ensemble. All that this requires is that the energy of a system be fixed. In fact, since the fireball is well-separated from the spectators, one may expect a mapping between the collisions and microcanonical ensemble to be good. However there are two obstructions, neither of them absolute. The first is that the microcanonical ensemble requires the energy of each member of the ensemble to be the same. This is hard to ensure without more control on centrality fluctuations than is possible at present. Secondly, one needs 4π-detectors to capture the entire energy of the fireball. Most detectors in use today miss a very large fraction of the energy.

So one must to try to map the ensemble of events recorded in the detector to either the canonical or a *grand-canonical ensemble*. The difference between these is that the system being studied must exchange either energy or material and energy (respectively) with a much larger system called a heat-bath. Since detectors accept particles from only a part of each fireball, one may be able to map the events on to a grand-canonical ensemble. Of course, thermal and chemical equilibrium is necessary in order to be able to do this.

We have already discussed the evidence that there is a degree of thermal and chemical equilibration at freeze out. So, if one observes a small part of the fireball, it may be possible to treat it in a grand-canonical ensemble where the rest of the fireball acts as the heat-bath. In order to make sure that the system (observed fraction of the fireball) is much smaller than the heat-bath (the unobserved fraction), one should use as small an angular coverage as possible while keeping the observed volume much larger than any intrinsic correlation volumes in the fireball (see Figure 14). If the acceptance in rapidity is Δy,

V_s is the size of the system, and V_b is the volume of the heat-bath, then

$$\frac{V_s}{V_s + V_b} = \frac{\Delta y}{2\log(\sqrt{S}/M_p)}\,. \tag{13}$$

Taking $\Delta y = 2$, one sees that V_b/V_s is about 4.3 at the top RHIC energy of 200 GeV, and around 7.5 at $\sqrt{S} = 5$ TeV. These may be acceptable numbers. However, at $\sqrt{S} = 20$ GeV the ratio drops to 2, and by $\sqrt{S} = 5$ GeV, the "heat-bath" is smaller than the "system". In order to keep the ratio V_b/V_s fixed, one has to decrease Δy with the beam energy.

This may give rise to another problem, which is to keep the observed volume much larger than correlation lengths. If freezeout occurs at time τ_f, then the acceptance region, Δy, corresponds approximately to a distance $\Delta x = \tau_f \sinh(\Delta y)$. As long as correlation lengths are linear in the inverse freezeout temperature $1/T_f$, it is interesting to examine

$$\Delta x T_f = (\tau_f T_f) \sinh\left(\frac{2\log\sqrt{S}}{1 + V_b/V_s}\right). \tag{14}$$

If one wants $V_b/V_s \simeq 4$ at $\sqrt{S} = 5$, where $T_f \simeq 145$ MeV, and one takes $\tau_f \simeq 5$ fm, then one finds $\Delta x T_f \simeq 2.5$. This is a reasonable number, but it implies that $\Delta y = 0.65$ at this energy. Such a small acceptance window may cause statistics to drop significantly. However, for $\sqrt{S} \geq 20$ GeV, there is a good possibility that all these constraints may be satisfied simultaneously. Of course, if correlation lengths become very large at some $\sqrt{S}$ then all these arguments fail, and the system cannot be treated as being in equilibrium.

Conserved quantities, such as the net particle number or energy, can change by transport across the boundary of the system. As a result, energy and net particle numbers fluctuate in grand-canonical thermodynamics. These fluctuations can now be mapped into E/E fluctuations. They were first discussed and suggested as probes of the phase structure of QCD in [38]. The experimental variables which allow a direct comparison of QCD predictions with data were first discovered in [39], and the first lattice QCD predictions were made in [40].

The existence of fluctuations means that the baryon number or energy of an ensemble is not a fixed quantity, but has a probability distribution. Such a distribution is characterized by the cumulants, $[B^n]$, which are defined as the Taylor coefficients of the logarithm of the Laplace transform of the distribution, $P(B)$, of the baryon number—

$$\log\left[\int dB\, P(B)\, \mathrm{e}^{-sB}\right] = \sum_{n=1}^{\infty} [B^n] \frac{(-s)^n}{n!} \tag{15}$$

The cumulants are related to the Taylor coefficients of the expansion of the free energy [39] in terms of μ_B simply as

$$[B^n] = V_s \chi^{(n)}(T_f, \mu_B^f)\, T_f^{n-1}, \tag{16}$$

where $\chi^{(n)}(T, \mu_B)$ are generalized quark number susceptibilities ($\chi^{(1)}$ is the baryon density) [41]. As a result, ratios of the cumulants are independent of the factor V_s. These ratios depend on the ratios of the dimensionless quantities $\chi^{(n)}(T_f, \mu_B^f) T_f^{n-4}$, which can be computed in lattice QCD, as demonstrated in [40]. Similar ratios were also discussed in [42].

Since T_f and μ_B^f was already known from the analysis of yields, the experimental data could be compared to the lattice computation [43]. This comparison is reproduced in Figure 15. The remarkably good agreement has led to subsequent attempts to refine the comparison. These include following up the suggestions in [40] that the comparison could yield a measurement of T_c given T_f and μ_B^f [44] or the determination of T_f and μ_B^f given T_c [45]. There has also been a lot of work on various corrections

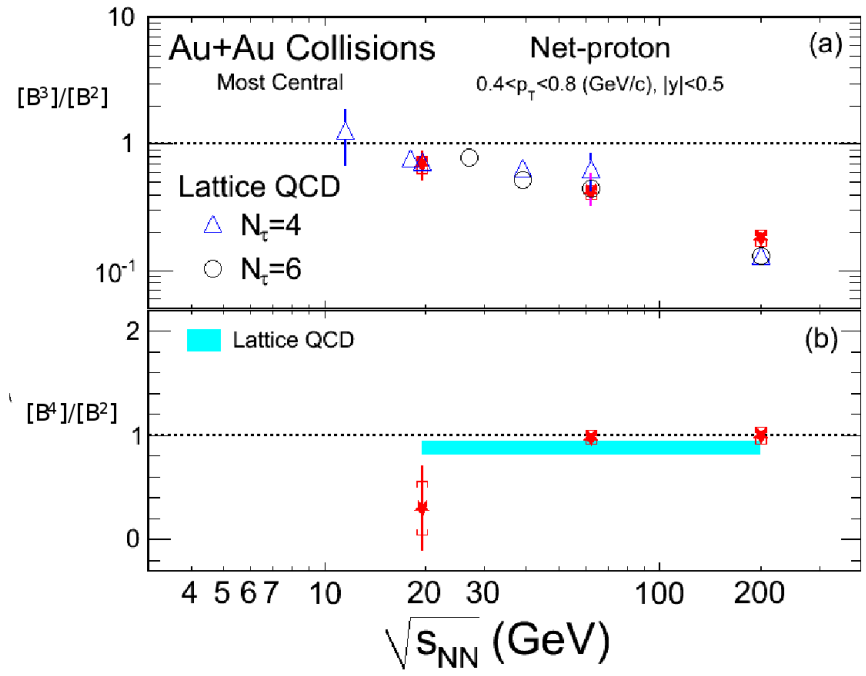


Fig. 15: A comparison of data on fluctuations of net proton number in the STAR experiment with the lattice QCD computations of [40] from [43].

which may need to be applied to the experimental data before comparing with predictions [46]. There is also sustained interest in fluid dynamical effects [29, 47]. In the meanwhile, much new experimental data has been added from the RHIC Beam Energy Scan (BES) [48]. A BES-II is expected shortly.

The long-term goal of the study of fluctuations is to understand the evolution of fluctuations along the freeze-out curve traced by changing the beam energy [39]. At higher energies one sees preliminary agreement between lattice predictions and experimental observations. This agreement is expected to break down in the vicinity of the QCD critical point because correlation lengths and relaxation times grow [49]. At lower energies one might expect a return to roughly thermal behaviour, although extracting this is fraught with theoretical and experimental challenges. The BES program aims to locate and study bulk matter near the QCD critical point.

8 Acknowledgement

I would like to thank the organizers of AEPSHEP 2014 for building a very stimulating scientific program, and for the wonderful local organization. It was a pleasure to lecture in this school. I would like to thank the Kavli Institute for Theoretical Physics China for hospitality during a part of the time that this manuscript was in preparation. I would like to thank Bedangadas Mohanty and Rishi Sharma for their helpful comments on the manuscript.

References

[1] S. Weinberg, *Quantum Theory of Fields: Modern Applications*, Vol II, Cambridge University Press (2010) Cambridge, Great Britain.

[2] R. D. Pisarski and F. Wilczek, *Phys. Rev.* D 29 (1984) 338; Y. Aoki *et al.*, *Phys. Lett.* B 643 (2006) 46 [hep-lat/0609068]; A. Bazavov *et al.*, *Phys. Rev.* D 85 (2012) 054503 [arxiv:1111.1710].

[3] J. Berges and K. Rajagopal, *Nucl. Phys.* B 538 (1999) 215 [hep-ph/9804233]; A. M. Halasz *et al.*, *Phys. Rev.* D 58 (1998) 096007 [hep-ph/9804290].

[4] D. T. Son and M. A. Stephanov, *Phys. Rev. Lett.* 86 (2001) 592 [hep-ph/0005225].

[5] R. V. Gavai and S. Gupta, *Phys. Rev.* D 66 (2002) 094510 [hep-lat/0208019]; S. Gupta, arxiv:0712.0434.

[6] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics: Statistical Physics*, Vol 5, Butterworth-Heinemann (1980) Oxford, Great Britain.

[7] Ph. de Forcrand, *PoS* LATTICE2009 (2009) 010 [arxiv:1005.0539]; S. Gupta, *PoS* LATTICE2010 (2010) 007 [arxiv:1101.0109]; L. Levkova, *PoS* LATTICE2011 (2011) 011 [arxiv:1201.1516]; G. Aarts, *PoS* LATTICE 2012 (2012) 017 [arxiv:1302.3028]; C. Gattringer, *PoS* LATTICE2013 (2014) 002 [arxiv:1401.7788].

[8] R. V. Gavai and S. Gupta, *Phys. Rev.* D 71 (2005) 114014 [hep-lat/0412035]; R. V. Gavai and S. Gupta, *Phys. Rev.* D 78 (2008) 114503 [arxiv:0806.2233]; S. Datta *et al.*, *PoS* LATTICE2013 (2014) 202; S. Gupta *et al.*, *Phys. Rev.* D 90 (2014) 034001 [arxiv:1405.2206].

[9] R. V. Gavai *et al.*, *Phys. Rev,* D 77 (2008) 114506 [arxiv:0803.0182]; A. Bazavov *et al.*, *Phys. Rev,* D 86 (2012) 094503 [arxiv:1205.3535]; C. Bonati *et al.*, *Phys. Rev. Lett.* 110 (2013) 252003 [arxiv:1301.7640]; G. Cossu *et al.*, *Phys. Rev.* D 87 (2013) 114514 [arxiv:1304.6145]; M. I. Buchoff *et al.*, *Phys. Rev.* D 89 (2014) 054514 [arxiv:1309.4149]; T.-W. Chiu *et al.*, *PoS* LATTICE2013 (2014) 165 [arxiv:1311.6220]; V. Dick *et al.*, *Phys. Rev.* D 91 (2015) 094504 [arxiv:1502.06190].

[10] M. D'Elia *et al.*, *Phys. Rev.* D 82 (2010) 051501 [arxiv:1005.5365]; G. S. Bali *et al.*, *J. H. E. P.* 1202 (2012) 044 [arxiv:1111.4956]; V. Skokov, *Phys. Rev.* D 85 (2012) 034026 [arxiv:1112.5137]; E.-M. Ilgenfritz *et al.Phys. Rev.* D 85 (2012) 114504 [arxiv:1203.3360]; F. Bruckmann *et al.*, *J. H. E. P.* 1304 (2013) 112 [arxiv:1303.3972]; L. Levkova and C. De Tar, *Phys. Rev. Lett.* D 112 (2014) 012002 [arxiv:1309.1142]; C. Bonati *et al.Phys. Rev.* D 89 (2014) 054506 [arxiv:1310.8656]; G. S. Bali *et al.*, *J. H. E. P.* 1408 (2014) 177 [arxiv:1406.0269];

[11] R. Baier *et al.*, *Phys. Lett.* B 502 (2001) 51 [hep-ph/0009237]; A. Dumitru *et al.*, *Phys. Rev.* D 75 (2007) 025016 [hep-ph/0604149]; P. Romatschke and R. Venugopalan, *Phys. Rev.* D 74 (2006) 045011; Y. V. Kovchegov and A. Taliotis, *Phys. Rev.* C 76 (2007) 014905 [arxiv:0705.1234]; D. A. Teaney, arxiv:0905.2433; V. Balasubramanian *et al.*, *Phys. Rev.* 84 (2011) 026010 [arxiv:1103.2683]; A. Kurkela and G. D. Moore, *J. H. E. P.* 1112 (2011) 044 [arxiv:1107:5050]; K. Fukushima and F. Gelis, *Nucl. Phys.* A 874 (2012) 108 [arxiv:1106.1396]; T. Epelbaum and F. Gelis, *Nucl. Phys.* A 872 (2011) 210 [arxiv:1107.0668]; J.-P. Blaizot *et al.*, *Nucl. Phys.* A 873 (2012) 68 [arxiv: 1107.5296]; J. Berges *et al.*, *Phys. Rev.* D 89 (2014) 7, 074011 [arxiv: 1303:5650].

[12] N. Armesto *et al.*, *J. Phys.* G 35 (2008) 054001 [arxiv:0711.0974]; B. Miller *et al.*, *Ann. Rev. Nucl. Part. Sci.* 62 (2012) 361 [arxiv:1202.3233].

[13] C. Gale *et al.*, *Phys. Rev. Lett.* 110 (2013) 012302 [arxiv:1209.6330].

[14] K. Aamodt *et al.*, *Phys. Rev. Lett.* 105 (2010) 252301 [arxiv:1011.3916].

[15] M. Gyulassy and M. Plümer, *Phys. Lett.* B 243 (1990) 432; X.-N. Wang and M. Gyulassy, *Phys. Rev. Lett.* 68 (1992) 1480.

[16] STAR Collaboration, *Phys. Rev. Lett.* 91 (2003) 072304 [arxiv:nucl-ex/0306024]

[17] S. Chatrchyan *et al.*(CMS Collaboration), *Phys. Rev. Lett.* 106 (2011) 212301 [arxiv:1102.5435].

[18] Yan-Jie Lee, QM 2014: https://indico.cern.ch/ event/ 219436/ session/ 3/ contribution/ 728/ material/ slides/ 1.pdf

[19] R. Baier *et al.*, *Phys. Lett.* B 345 (1995) 277 [hep-ph/9411409]; P. B. Arnold *et al.*, *J. H. E. P.* 0206

(2002) 030 [hep-ph/0204343]; G. Baym *et al.*, *Phys. Lett.* B 644 (2007) 48 [hep-ph/0604209]; S. Peigne and A. V. Smilga, *Phys. Usp.* 52 (2009) 659 [arxiv:0810.5702]; S. Caron-Huot, *Phys. Rev.* D 79 (2009) 065039 [arxiv:0811.1603]. J. Ghiglieri and D. Teaney, arxiv:1502.03730.

[20] A. Majumder, *Phys. Rev.* C 87 (2013) 034905 [arxiv:1202.5295]; X. Ji, *Phys. Rev. Lett.* 110 (2013) 262002 [arxiv:1305.1539]; M. Panero *et al.*, *Phys. Rev. Lett.* 112 (2014) 162001 [arxiv:1307.5850]; M. Laine and A. Rothkopf, *PoS* LATTICE2013 (2013) 174 [arxiv:1310.2413].

[21] H. Liu *et al.*, *Phys. Rev. Lett.* 98 (2007) 182301 [hep-ph/0607062]; H. Liu *et al.*, *J. H. E. P.* 0703 (2007) 066 [hep-ph/0612168]; J. Casalderrey-Solana and D. Teaney, *J. H. E. P.* 0704 (2007) 039 [hep-th/0701123]; Y. Hatta, E. Iancu and A. H. Mueller *J. H. E. P.* 0805 (2008) 037 [arxiv:0803.2481]; P. M. Chesler *et al.*, *Phys. Rev.* D 79 (2009) 125015 [arxiv:0810.1985]; J. de Boer *et al.*, *J. H. E. P.* 0907 (2009) 094 [arxiv:0812.5112].

[22] J. D. Bjorken, *Phys. Rev.* D 27 (1983) 140.

[23] J.-Y. Ollitrault, *Phys. Rev.* D 46 (1992) 229.

[24] S. Voloshin and Y. Zhang, *Z. Phys.* C 70 (1996) 665 [hep-ph/9407282].

[25] A. P. Mishra *et al.*, *Phys. Rev.* C 81 (2010) 034903 [arxiv:0811.0292]; B. Alver and G. Roland, *Phys. Rev.* C 81 (2010) 054905 [arxiv:1003.0194]; D. Teaney and Li Yan, *Phys. Rev.* C 83 (2011) 064904 [arxiv:1010.1876].

[26] H.J. Dresher *et al.*, *Phys. Rev./* C 76 (2007) 024905 [arxiv:0704.3553]; P. Romatschke and U. Romatschke, *Phys. Rev. Lett.* 99 (2007) 172301 [arxiv:0706.1522]; H. Song and U. W. Heinz, *Phys. Lett.* B 658 (2008) 279 [arxiv:0709.0742]. K. Dusling and D. Teaney, *Phys. Rev.* C 77 (2008) 034905 [arxiv:0710.5932]

[27] M. Luzum and P. Romatschke, *Phys. Rev.* C 78 (2008) 034915 [arxiv:0804.4015]; G. Ferini *et al.*, *Phys. Lett.* B 670 (2009) 325 [arxiv:0805.4814]; T. Hirano and Y. Nara, *Phys. Rev.* C 79 (2009) 064904 [arxiv:0904.4080]; K. Dusling *et al.*, *Phys. Rev.* C 81 (2010) 034907 [arxiv:0909.0754]; B. Schenke *et al.*, *Phys. Rev.* C 82 (2010) 014903 [arxiv:1004.1408]; H. Holopainen *et al.*, *Phys. Rev.* C 83 (2011) 034901 [arxiv:1007.0368]; G.Y. Qin *et al.*, *Phys. Rev.* C 82 (2010) 064903 [arxiv:1009.1847]; C. Shen *et al.*, *Phys. Rev.* C 84 (2011) 044903 [arxiv:1105.3226]; P. Bozek, *Phys. Rev.* C 85 (2012) 034901 [arxiv:1110.6742]; M. Martinez *et al.*, *Phys. Rev.* C 85 (2012) 064913 [arxiv:1204.1473]; D. Teaney and Li Yan, *Phys. Rev.* C 86 (2012) 044908 [arxiv:1206.1905]; STAR Collaboration, *Phys. Rev.* C 88 (2013) 014904 [arxiv:1301.2187]; CMS Collaboration, *Phys. Lett.* B 724 (2013) 213 [arxiv:1305.0609]; ATLAS Collaboration, *J. H. E. P.* 1311 (2013) 183 [arxiv:1305.2942]; ALICE Collaboration, *Phys. Lett.* B 726 (2013) 164 [arxiv:1307.3237].

[28] J. Cleymans and K. Redlich, *Phys. Rev. Lett.* 81 (1998) 5284 [nucl-th/9808030]; Z.-W. Lin *et al.*, *Phys. Rev.* C 72 (2005) 064901 [nucl-th/064901]; A. Andronic *et al.*, *Nucl. Phys.* A 772 (2006) 167 [nucl-th/0511071]; F. Becattini *et al.*, *Phys. Rev.* C73 (2006) 044905 [hep-ph/0511092]; NA49 Collaboration *Phys. Rev.* C 73 (2006) 044910; V. V. Begun *et al.*, *Phys. Rev.* C 76 (2007) 024902 [nucl-th/0611075]; F. Becattini and J. Manninen, *J. Phys.* G 35 (2008) 104013 [arxiv:0805.0098]; STAR Collaboration, arxiv:1007.2613; M. Chojnacki *et al.*, *Comput. Phys. Commun.* 183 (2012) 746 [arxiv:1102.0273]; ALICE Collaboration, *Phys. Rev.* C 88 (2013) 044910 [arxiv:1303.0737]; S. Borsanyi *et al.*, *Phys. Rev. Lett.* 111 (2013) 062005 [arxiv:1305.5161].

[29] R. S. Bhalerao and S. Gupta, *Phys. Rev.* C 79 (2009) 064901 [arxiv:0901.4677].

[30] S. Chatterjee *et al.*, *Phys. Lett.* B 727 (2013) 554 [arxiv:1306.2006]; K. A. Bugaev *et al.*, *Europhys. Lett.* 104 (2013) 22002 [arxiv:1308.3594].

[31] J. Steinheimer *et al.*, *Phys. Rev. Lett.* 110 (2013) 042501 [arxiv:1203.5302]; F. Becattini *et al.Phys. Rev. Lett.* 111 (2013) 082302 [arxiv:1212.2431].

[32] T. Matsui and H. Satz, *Phys. Lett.* B 178 (1986) 416.

[33] M. Asakawa and T. Hatsuda, *Phys. Rev. Lett.* 92 (2004) 012001 [hep-lat/0308034]; S. Datta *et al.*,

Phys. Rev. D 69 (2004) 094507 [hep-lat/0312037].

[34] F. Karsch *et al.*, *Phys. Lett.* B 637 (2006) 75 [hep-ph/0512239]; H. Satz, *Nucl. Phys.* A 783 (2007) 249.

[35] S. Chatrchyan *et al.Phys. Rev. Lett.* 109 (2012) 222301 [arxiv:1208.2826]; S. Chatrchyan *et al.Phys. Rev. Lett.* 107 (2011) 052302 [arxiv:1105.4894].

[36] S. Gupta and R. Sharma, *Phys. Rev.* C 89 (2014) 057901 [arxiv:1401.2930].

[37] N. Brambilla *et al.*, *Eur. Phys. J.* C 71 (2011) 1534, and references therein.

[38] M. A. Stephanov *et al.*, *Phys. Rev. Lett.* 81 (1998) 4816 [hep-ph/9806219]; M. A. Stephanov *et al.*, *Phys. Rev.* D 60 (1999) 114028 [hep-ph/9903292].

[39] S. Gupta, *PoS* CPOD2009 (2009) 025 [arxiv:0909.4630]; S. Gupta, *Prog. Theor. Phys. Suppl.* 186 (2010) 440.

[40] R. V. Gavai and S. Gupta, *Phys. Lett.* B 696 (2011) 459 [arxiv:1001.3796].

[41] R. V. Gavai and S. Gupta, *Phys. Rev.* D 68 (2003) 034506 [hep-lat/0303013].

[42] C. Athanasiou *et al.*, *Phys. Rev.* D 82 (2010) 074008 [arxiv:1006.4636].

[43] M. M. Aggarwal *et al.*(STAR Collaboration), *Phys. Rev. Lett.* 105 (2010) 022302 [arxiv:1004.4959].

[44] S. Gupta *et al.*, *Science* 332 (2011) 1525 [arxiv:1105.3934].

[45] A. Bazavov *et al.*, *Phys. Rev. Lett.* 109 (2012) 192302 [arxiv:1208.1220]; S. Borsanyi *et al.*, *Phys. Rev. Lett.* 111 (2013) 062005 [arxiv:1305.5161].

[46] E. S. Fraga *et al.*, *Phys. Rev.* C 84 (2011) 011903 [arxiv:1104.3755]; M. Kitazawa and M. Asakawa *Phys. Rev.* C 85 (2012) 021901 [arxiv:1107.1412]; A. Bzdak etal, *Phys. Rev.* C 87 (2013) 014901 [arxiv:1203.4529]; M. Kitazawa and M. Asakawa *Phys. Rev.* C 86 (2012) 024904 [arxiv:1205.3292]; V. Skokov and B. Friman, *Phys. Rev.* C 88 (2013) 034911 [arxiv:1205.4756]; A. Bzdak and V. Koch, *Phys. Rev.* C 86 (2012) 044904 [arxiv:1206.4286]; X. Luo *et al.*, *J. Phys.* G 40 (2013) 105104 [arxiv:1302.2332]; H. Ono *et al.*, *Phys. Rev.* C 87 (2013) 041901 [arxiv:1303.3338]; P. Garg *et al.*, *Phys. Lett.* B 726 (2013) 691 [arxiv:1304.7133]; A. Tang and G. Wang, *Phys. Rev.* C 88 (2013) 024905 [arxiv:1305.1392]; L. Chen *et al.*, *J. Phys.* G 41 (2014) 105107 [arxiv:1312.0749]; A. Bzdak and V. Koch, *Phys. Rev.* C 91 (2015) 027901 [arxiv:1312.4574]; X. Luo *et al.*, *Nucl. Phys.* A 931 (2014) 808 [arxiv:1408.0495]; M. Sakaida *et al.*, *Phys. Rev.* C 90 2014) 064911 [arxiv:1409.6866].

[47] M. A. Stephanov, *Phys. Rev.* D 81 (2010) 054012 [arxiv:0911.1772]; K. Xiao *et al.*, *Chin. Phys.* C 35 (2011) 467; J. I. Kapusta and J. M. Torres-Rincon, *Phys. Rev.* C 86 (2012) 054911 [arxiv:1209.0675]; M. Kitazawa, arxiv: 1505.04349; S. Mukherjee *et al.*, arxiv:1506.00645

[48] X. Zhang (STAR Collaboration) *Nucl. Phys.* A 904-905 (2013) 543c; A. Sarkar (STAR Collaboration) *PoS* CPOD2013 (2013) 043; L. Adamczyk *et al.*(STAR Collaboration) *Phys. Rev. Lett.* 112 (2014) 032302 [arxiv:1309.5681]; L. Adamczyk *et al.*(STAR Collaboration) *Phys. Rev. Lett.* 113 (2014) 092301 [arxiv:1402.1558]; A. Adare *et al.*(PHENIX Collaboration) arxiv:1506.07834.

[49] B. Berdnikov and K. Rajagopal, *Phys. Rev.* D 61 (2000) 105017 [hep-ph/9912274]; M. A. Stephanov, *Phys. Rev. Lett.* 102 (2009) 032301 [arxiv:0809.3450].

Proceedings of the 2014 Asia–Europe–Pacific School of High-Energy Physics, Puri, India, 4–17 November 2014, edited by
M. Mulders and R. Godbole, CERN Yellow Reports: School Proceedings, Vol. 2/2017, CERN-2017-005-SP (CERN, Geneva, 2017)

Cosmology

V.A. Rubakov
Institute for Nuclear Research of the Russian Academy of Sciences,
Moscow, Russia
and
Department of Particle Physics and Cosmology, Physics Faculty, Moscow State University,
Moscow, Russia

Abstract
Cosmology and particle physics are deeply interrelated. Among the common problems are dark energy, dark matter and baryon asymmetry of the Universe. We discuss these problems in general terms, and concentrate on several particular hypotheses. On the dark matter side, we consider weakly interacting massive particles and axions/axion-like particles as cold dark matter, sterile neutrinos and gravitinos as warm dark matter. On the baryon asymmetry side, we discuss electroweak baryogenesis as a still-viable mechanism. We briefly describe diverse experimental and observational approaches towards checking these hypotheses. We then turn to the earliest cosmology. We give arguments showing that the hot stage was preceded by another epoch at which density perturbations and possibly primordial gravity waves were generated. The best guess here is inflation, which is consistent with everything we know of density perturbations, but there are alternative scenarios. Future measurements of the properties of density perturbations and possible discovery of primordial gravity waves have strong potential in this regard.

Keywords
Lectures; cosmology; cosmological model; baryon asymmetry; dark matter; dark energy; nucleosynthesis.

1 Introduction

Cosmology is one of the major sources of inspiration—and confusion—for particle physicists. It gives direct evidence for the necessity to extend the Standard Model of particle physics, possibly at an energy scale that can be probed by collider experiments. Indeed, there is no doubt that most part of the mass in the present Universe is in the form of mysterious dark matter particles which are not present in the Standard Model. Also, the very existence of conventional matter in our Universe (i.e., matter–antimatter asymmetry) calls for processes with baryon number violation and substantial charge parity (CP)-violation, which have not been observed in experiments. These processes had to be rapid in the early Universe and, furthermore, the asymmetry between matter and antimatter had to be generated in a fairly turbulent cosmological epoch. Again, the conditions necessary for the generation of this asymmetry are not present in the Standard Model. Solving the problems of dark matter and matter–antimatter asymmetry are the two immediate challenges for particle physics.

Going very much back into the cosmological history, we encounter another challenging issue. It is very well known that matter in the Universe was very hot and dense early on. It is less known that the properties of the matter distribution in the past and present Universe, reflected in the properties of the cosmic microwave background (CMB), galaxy distribution etc, unambiguously tell us that the hot epoch was not the earliest. It was preceded by another, completely different epoch responsible for the generation of inhomogeneities which in the end have become galaxies and their clusters, stars and ourselves. Obviously, the very fact that we are confident about the existence of such an epoch

https://doi.org/10.23730/CYRSP-2017-002.239

is a fundamental result of theoretical and observational cosmology. The most plausible hypothesis on that epoch is cosmological inflation, though the observational support of this scenario is presently not overwhelming, and alternative possibilities have not been ruled out. For the time being it appears unlikely that we will be able to probe the physics behind that epoch in terrestrial experiments, but there is no doubt that this physics belongs to the broad domain of 'particles and fields'.

After this brief introduction, the scope of these lectures must be clear. To set the stage, we briefly consider the basic notions of cosmology. We then discuss several dark matter particle candidates and mechanisms for dark matter generation. Needless to say, these candidates do not exhaust the long list of the candidates proposed; our choice is based on a personal view of what candidates are more plausible. Our next topic is the matter–antimatter asymmetry of the Universe, and we present electroweak baryogenesis as a mechanism particularly interesting from the viewpoint of the LHC experiments. The last part of these lectures deals with cosmological perturbations, inflation (and its alternatives) and the potential of future observational data.

These lectures are meant to be self-contained, but we necessarily omit numerous details, while trying to make clear the basic ideas and results. More complete accounts of cosmology and its particle-physics aspects may be found in various books [1–6]. Dark matter candidates we consider in these lectures are reviewed in Refs. [7–10]. Electroweak baryogenesis is presented in detail in reviews [11–13]; for reference, a plausible alternative scenario, leptogenesis, is discussed in reviews [14, 15]. Aspects of inflation and its alternatives are reviewed in Refs. [16–20].

2 Expanding universe

2.1 Friedmann–Lemaître–Robertson–Walker metric

Our Universe (more precisely, its visible part) is *homogeneous and isotropic*. Clearly, this does not apply to relatively small spatial scales: there are galaxies, clusters of galaxies and giant voids. But boxes of sizes exceeding about 200 Mpc all look the same. Here the Mpc is the distance unit conventionally used in cosmology,

$$1 \text{ Mpc} \approx 3 \times 10^6 \text{ light years} \approx 3 \times 10^{24} \text{ cm} .$$

There are three types of homogeneous and isotropic three-dimensional spaces, labelled by an integer parameter $\varkappa$. These are three-sphere (closed model, $\varkappa = +1$), flat (Euclidean) space (flat model, $\varkappa = 0$) and three-hyperboloid (open model, $\varkappa = -1$). We will see that the parameter $\varkappa$ enters the dynamical equations governing the space–time fabric of the Universe.

Another basic property of our Universe is that it *expands*. This is encoded in the space–time metric

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(t)\mathrm{d}\mathbf{x}^2 , \tag{1}$$

where $\mathrm{d}\mathbf{x}^2$ is the distance on a unit three-sphere, Euclidean space or hyperboloid. The metric (1) is called the Friedmann–Lemaître–Robertson–Walker (FLRW) metric, and $a(t)$ is the scale factor. In these lectures we use natural units, setting the speed of light and Planck and Boltzmann constants equal to 1,

$$c = \hbar = k_{\mathrm{B}} = 1 .$$

In these units, Newton's gravity constant is $G = M_{\mathrm{Pl}}^{-2}$, where $M_{\mathrm{Pl}} = 1.2 \times 10^{19}$ GeV is the Planck mass.

The meaning of Eq. (1) is as follows. One can check that a free mass put at a certain $\mathbf{x}$ at zero velocity will stay at the same $\mathbf{x}$ forever. In other words, the coordinates $\mathbf{x}$ are comoving. The scale factor $a(t)$ increases in time, so the distance between free masses of fixed spatial coordinates $\mathbf{x}$ grows, $\mathrm{d}l^2 = a^2(t)\mathrm{d}\mathbf{x}^2$. The space stretches out; the galaxies run away from each other.

This expansion manifests itself as a red shift. Red shift is often interpreted as the Doppler effect for a source running away from us with velocity v: if the wavelength at emission is λ_{e}, then the wavelength

we measure is $\lambda_0 = (1+z)\lambda_{\rm e}$, where $z = v/c$ (here we temporarily restore the speed of light). This interpretation is useless and rather misleading in cosmology (with respect to which reference frame does the source move?). The correct interpretation is that as the Universe expands, space stretches out and the photon wavelength increases proportionally to the scale factor a. So, the relation between the wavelengths is

$$\lambda_0 = (1+z)\lambda_{\rm e} \ , \quad \text{where } z = \frac{a(t_0)}{a(t_{\rm e})} - 1 \ ,$$

where $t_{\rm e}$ is the emission time. For $z \ll 1$, this relation reduces to the Hubble law,

$$z = H_0 r \ , \tag{2}$$

where r is the physical distance to the source and $H_0 \equiv H(t_0)$ is the present value of the Hubble parameter

$$H(t) = \frac{\dot{a}(t)}{a(t)} \ .$$

In the formulas above, we label the present values of time-dependent quantities by subscript 0; we will always do so in these lectures.

Question. Derive the Hubble law (2) for $z \ll 1$.

The red shift of an object is directly measurable. The wavelength $\lambda_{\rm e}$ is fixed by physics of the source, say, it is the wavelength of a photon emitted by an excited hydrogen atom. So, one identifies a series of emission or absorption lines, thus determining $\lambda_{\rm e}$, and measures their actual wavelengths λ_0. These spectroscopic measurements give accurate values of z even for distant sources. On the other hand, the red shift is related to the time of emission and hence to the distance to the source. Absolute distances to astrophysical sources have a lot more systematic uncertainty, and so do the direct measurements of the Hubble parameter H_0. According to the Planck Collaboration [21], the combination of observational data gives

$$H_0 = (67.8 \pm 0.9)\ \frac{\rm km}{\rm s\ Mpc} \approx (14.4 \times 10^9\ {\rm yr})^{-1} \ , \tag{3}$$

where the unit used in the first expression reflects the interpretation of red shift in terms of the Doppler shift. The fact that the systematic uncertainties in the determination of H_0 are pretty large is illustrated in Fig. 1.

Traditionally, the present value of the Hubble parameter is written as

$$H_0 = h \times 100\ \frac{\rm km}{\rm s\ Mpc} \ . \tag{4}$$

Thus, $h \approx 0.7$. We will use this value in further estimates.

2.2 Hot Universe: recombination, Big Bang nucleosynthesis and neutrinos

Our Universe is filled with CMB. The CMB as observed today consists of photons with an excellent black-body spectrum of temperature

$$T_0 = 2.7255 \pm 0.0006\ {\rm K} \ . \tag{5}$$

The spectrum has been precisely measured by various instruments, see Fig. 2, and does not show any deviation from the Planck spectrum (see Ref. [23] for a detailed review).

Once the present photon temperature is known, the number density and energy density of CMB photons are known from the Planck distribution formulas,

$$n_{\gamma,0} = 410\ \ {\rm cm}^{-3} \ , \qquad \rho_{\gamma,0} = \frac{\pi^2}{15} T_0^4 = 2.7 \times 10^{-10}\ \frac{\rm GeV}{\rm cm^3} \tag{6}$$

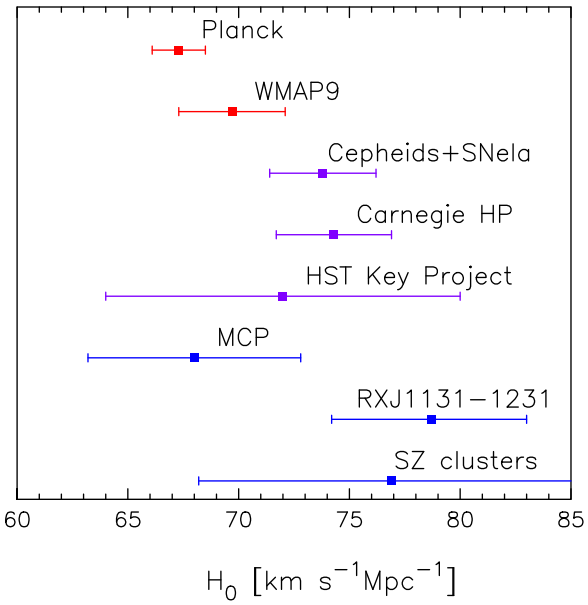


Fig. 1: Recent determinations of the Hubble parameter H_0 [22]

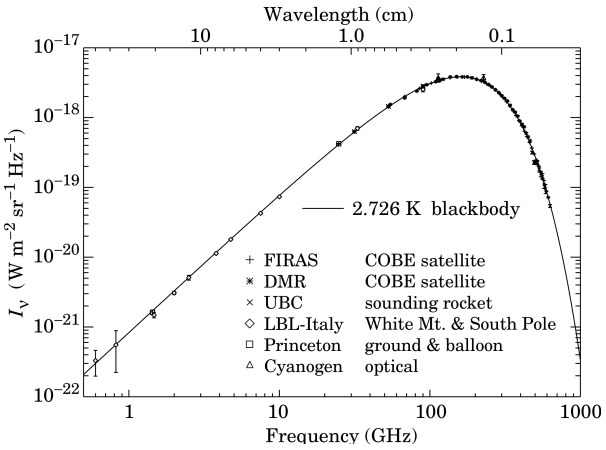


Fig. 2: Measured CMB energy spectrum as compiled in Ref. [24]

(the second expression is the Stefan–Boltzmann formula).

The CMB is a remnant of an earlier cosmological epoch. The Universe was hot at early times and, as it expands, the matter in it cools down. Since the wavelength of a photon evolves in time as $a(t)$, its energy and hence temperature scale as

$$\omega(t) \propto a^{-1}(t) \,, \qquad T(t) = \frac{a_0}{a(t)} T_0 = (1+z) T_0 \,.$$

When the Universe was hot, the usual matter (electrons and protons with a rather small admixture of light nuclei, mainly ^{4}He) was in the plasma phase. At that time photons strongly interacted with electrons due to the Thomson scattering and protons interacted with electrons via the Coulomb force, so all these particles were in thermal equilibrium. As the Universe cooled down, electrons 'recombined' with protons into neutral hydrogen atoms (helium recombined earlier), and the Universe became transparent to photons: at that time, the density of hydrogen atoms was quite small, 250 cm^{-3}. The photon last

scattering occurred at temperature and red shift

$$T_{\rm rec} \approx 3000 \text{ K} , \qquad z_{\rm rec} \approx 1090 ,$$

when the age of the Universe was about $t \approx 380$ thousand years (for comparison, its present age is about 13.8 billion years). Needless to say, CMB photons got red shifted since the last scattering, so their present temperature is $T_0 = T_{\rm rec}/(1 + z_{\rm rec})$.

The photon last scattering epoch is an important cornerstone in the cosmological history. Since after that CMB photons travel freely through the Universe, they give us a photographic picture of the Universe at that epoch. Importantly, the duration of the last scattering epoch was considerably shorter than the Hubble time $H^{-1}(t_{\rm rec})$; to a reasonable approximation, recombination occurred instantaneously. Thus, the photographic picture is only slightly washed out due to the finite thickness of the last scattering surface.

At even earlier times, the temperature of the Universe was even higher. We have direct evidence that at some point the temperature in the Universe was in the MeV range. A traditional source of evidence is the Big Bang nucleosynthesis (BBN). The story begins at a temperature of about 1 MeV, when the age of the Universe was about 1 s. Before that time neutrons were rapidly created and destroyed in weak processes like

$$\mathrm{e}^- + \mathrm{p} \longleftrightarrow \mathrm{n} + \nu_{\rm e} , \tag{7}$$

while at $T_{\rm n} \approx 1$ MeV these processes switched off, and the comoving number density of neutrons froze out. The neutron-to-proton ratio at that time was given by the Boltzmann factor,

$$\frac{n_{\rm e}}{n_{\rm p}} = \mathrm{e}^{-\frac{m_{\rm n}-m_{\rm p}}{T_{\rm n}}} .$$

Interestingly, $m_{\rm n} - m_{\rm p} \sim T_{\rm n}$, so the neutron–proton ratio at neutron freeze-out and later was neither equal to 1, nor very small. Were it equal to 1, protons would combine with neutrons into ^{4}He at a somewhat later time, and there would remain no hydrogen in the Universe. On the other hand, for very small $n_{\rm n}/n_{\rm p}$, too few light nuclei would be formed, and we would not have any observable remnants of the BBN epoch. In either case, the Universe would be quite different from what it actually is. It is worth noting that the approximate relation $m_{\rm n} - m_{\rm p} \sim T_{\rm n}$ is a coincidence: $m_{\rm n} - m_{\rm p}$ is determined by light quark masses and electromagnetic coupling, while $T_{\rm n}$ is determined by the strength of weak interactions (which govern the rates of the processes (7)) and gravity (which governs the expansion of the Universe). This is one of numerous coincidences we encounter in cosmology.

At temperatures somewhat below $T_{\rm n}$, the neutrons combined with protons into light elements in thermonuclear reactions like

$$\begin{aligned} \mathrm{p} + \mathrm{n} &\to {}^2\mathrm{H} + \gamma , \\ {}^2\mathrm{H} + \mathrm{p} &\to {}^3\mathrm{He} + \gamma , \\ {}^3\mathrm{He} + {}^2\mathrm{H} &\to {}^4\mathrm{He} + \mathrm{p} , \end{aligned} \tag{8}$$

etc, up to ^{7}Li. The abundances of light elements have been measured; see Fig. 3. On the other hand, the only parameter relevant for calculating these abundances (assuming negligible neutrino–antineutrino asymmetry) is the baryon-to-photon ratio

$$\eta_{\rm B} \equiv \eta = \frac{n_{\rm B}}{n_\gamma} , \tag{9}$$

characterizing the number density of baryons. Comparison of the BBN theory with the observational determination of the composition of the cosmic medium enables one to determine $\eta_{\rm B}$ and check the overall consistency of the BBN picture. It is even more reassuring that a completely independent measurement

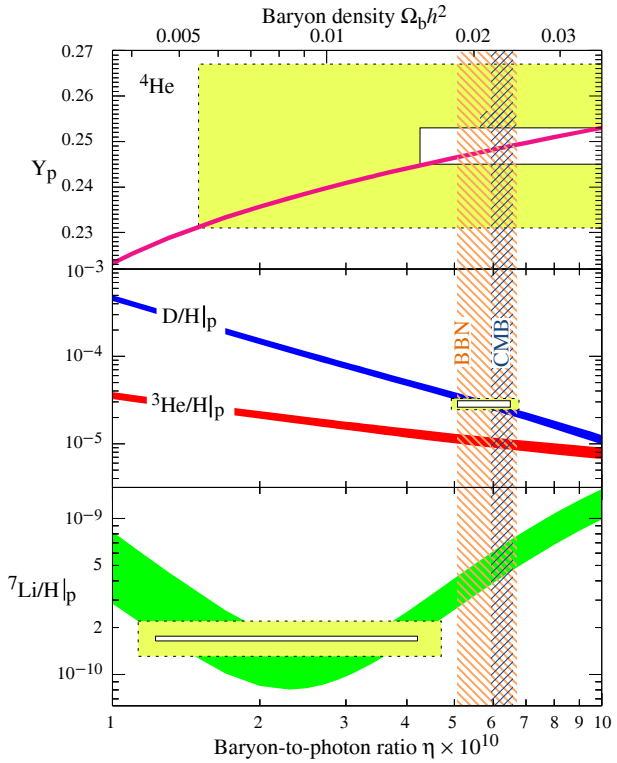


Fig. 3: Abundances of light elements, measured (boxes; larger boxes include systematic uncertainties) and calculated as functions of baryon-to-photon ratio η [25]. The determination of $\eta \equiv \eta_{\rm B}$ from BBN (vertical range marked BBN) is in excellent agreement with the determination from the analysis of CMB temperature fluctuations (vertical range marked CMB).

of $\eta_{\rm B}$ that makes use of the CMB temperature fluctuations is in excellent agreement with BBN. Thus, BBN gives us confidence that we understand the Universe at $T \sim 1$ MeV, $t \sim 1$ s. In particular, we are convinced that the cosmological expansion was governed by general relativity.

Another class of processes of interest at temperatures in the MeV range is neutrino production, annihilation and scattering,

$$\nu_\alpha + \bar{\nu}_\alpha \longleftrightarrow {\rm e}^+ + {\rm e}^-$$

and crossing processes. Here the subscript α labels neutrino flavours. These processes switch off at $T \sim$ 2–3 MeV, depending on neutrino flavour. Since then neutrinos do not interact with the cosmic medium other than gravitationally, but they do affect the properties of CMB and distribution of galaxies through their gravitational interactions. These effects are not negligible, since the energy density of relativistic neutrinos is almost the same as that of photons and, at temperature $T_{\rm rec} \simeq 3000$ K, the energy density of these relativistic species is only three times smaller than the energy density of non-relativistic particles (dark matter and baryons). Thus, observational data can be used to establish, albeit somewhat indirectly, the existence of relic neutrinos and set limits on neutrino masses. An example is shown in Fig. 4, where the number of neutrino flavours $N_{\rm eff}$ and the sum of neutrino masses are taken as free parameters. We see that cosmology *requires* relic neutrinos of at least three flavours and sets the limit on neutrino mass $m_\nu \lesssim 0.1$ eV (neutrino oscillation data tell that neutrinos with masses above 0.1 MeV are degenerate in mass). The latest Planck analysis gives [21]

$$\sum_i m_{\nu_i} < 0.23 \text{ eV} , \qquad N_{\rm eff} = 3.15 \pm 0.23 .$$

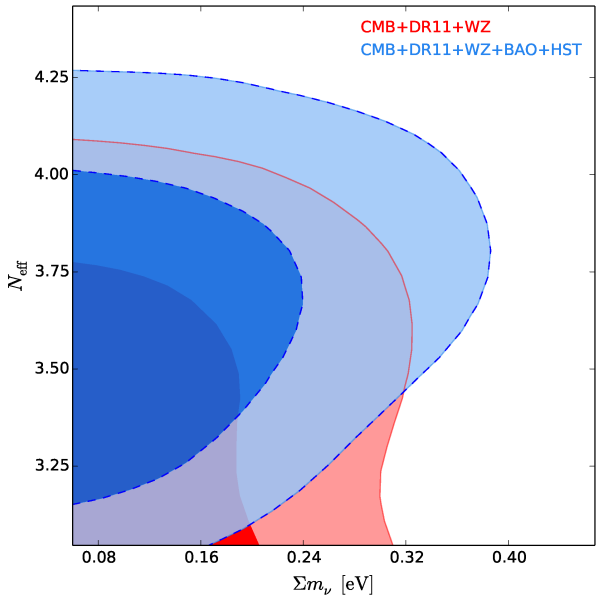


Fig. 4: Effective number of neutrino species and sum of neutrino masses allowed by cosmological observations [26].

2.3 Dynamics of expansion

The basic equation governing the expansion rate of the Universe is the Friedmann equation,

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi}{3M_{\rm Pl}^2}\rho - \frac{\varkappa}{a^2}\,, \tag{10}$$

where the dot denotes derivative with respect to time t, ρ is the *total* energy density in the Universe and $\varkappa = 0, \pm 1$ is the parameter, introduced in Section 2.1, that discriminates the Euclidean 3-space ($\varkappa = 0$) and curved 3-spaces. The Friedmann equation is nothing but the (00)-component of the Einstein equations of general relativity, $R_{00} - \frac{1}{2}g_{00}R = 8\pi T_{00}$, specified to the FLRW metric. Observationally, the spatial curvature of the Universe is very small: the last, curvature term in the right-hand side of Eq. (10) is small compared to the energy density term [21],

$$\frac{1/a^2}{8\pi\rho/(3M_{\rm Pl}^2)} < 0.005\,,$$

while the theoretical expectation is that the spatial curvature is completely negligible. Establishing that the three-dimensional space is (nearly) Euclidean is one of the profound results of CMB observations.

In what follows we set $\varkappa = 0$ and write the Friedmann equation as

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi}{3M_{\rm Pl}^2}\rho\,. \tag{11}$$

The standard parameter used in cosmology is the critical density,

$$\rho_{\rm c} = \frac{3}{8\pi}M_{\rm Pl}^2 H_0^2 \approx 5\times 10^{-6}\,\frac{\rm GeV}{\rm cm^3}\,. \tag{12}$$

According to Eq. (11), it is equal to the sum of all forms of energy density in the present Universe. There are at least three of such forms: relativistic matter, or *radiation*, non-relativistic *matter, M* and

dark energy, Λ. For every form λ with the *present* energy density $\rho_{\lambda,0}$, one defines the parameter

$$\Omega_\lambda = \frac{\rho_{\lambda,0}}{\rho_{\rm c}} \ .$$

One finds from Eq. (11) that

$$\sum_\lambda \Omega_\lambda = 1 \ .$$

The Ω are important cosmological parameters characterizing the energy balance in the present Universe. Their numerical values are

$$\Omega_{\rm rad} = 8.7 \times 10^{-5} \ , \tag{13a}$$

$$\Omega_{\rm M} = 0.31 \ , \tag{13b}$$

$$\Omega_\Lambda = 0.69 \ . \tag{13c}$$

The value of $\Omega_{\rm rad}$ needs qualification. At early times, when the temperature exceeds the masses of all neutrino species, neutrinos are relativistic. The value of $\Omega_{\rm rad}$ in Eq. (13a) is calculated for the unrealistic case in which *all neutrinos are relativistic today*, so the radiation component even at present consists of CMB photons and three neutrino species. This prescription is convenient for studying the energy (and entropy) content in the early Universe, since it enables one to scale the energy density (and entropy) back in time in a simple way, see below. For future reference, let us give the value of the present entropy density in the Universe, pretending that neutrinos are relativistic,

$$s_0 \approx 3000 \ {\rm cm}^{-3} \ . \tag{14}$$

Question. Calculate the numerical value of Ω_γ and the entropy density of CMB photons.

Non-relativistic matter consists of baryons and dark matter. The contributions of each of these fractions are [21]

$$\Omega_{\rm B} = 0.048 \ ,$$
$$\Omega_{\rm DM} = 0.26 \ .$$

Different components of the energy density evolve differently in time. The energy of a given photon or massless neutrino scales as a^{-1}, and the number density of these species scales as a^{-3}. Therefore, the energy density of radiation scales as $\rho_{\rm rad} \propto a^{-4}$ and

$$\rho_{\rm rad}(t) = \left(\frac{a(t)}{a_0}\right)^4 \rho_{\rm rad,0} = (1+z)^4 \, \Omega_{\rm rad} \rho_{\rm c} \ . \tag{15}$$

The energy of non-relativistic matter is dominated by the mass of its particles, so the energy density scales as the number density, i.e.,

$$\rho_{\rm M}(t) = \left(\frac{a(t)}{a_0}\right)^3 \rho_{\rm M,0} = (1+z)^3 \, \Omega_{\rm M} \rho_{\rm c} \ . \tag{16}$$

Finally, the energy density of dark energy does not change in time, or changes very slowly. We assume for definiteness that ρ_Λ stays constant in time,

$$\rho_\Lambda = \Omega_\Lambda \rho_{\rm c} = {\rm const} \ . \tag{17}$$

In fact, whether or not ρ_Λ depends on time (even slightly) is a very important question. If dark energy is a cosmological constant (or, equivalently, vacuum energy), then it does not depend on time at all. Even

a slight dependence of ρ_Λ on time would mean that we are dealing with something different from the cosmological constant, like, e.g., a new scalar field with a very flat scalar potential. The existing limits on the time evolution of dark energy correspond, roughly speaking, to the variation of ρ_Λ by not more than 20% in the last 8 billion years (from the time corresponding to $z \approx 1$); usually these limits are expressed in terms of the equation-of-state parameter relating energy density and effective pressure $p_\Lambda = w_\Lambda \rho_\Lambda$:

$$w_\Lambda \approx 1.0 \pm 0.1 \; . \tag{18}$$

The relevance of the effective pressure is seen from the covariant conservation equation for the energy–momentum tensor, $\nabla_\mu T^{\mu\nu} = 0$, whose $\nu = 0$ component reads

$$\dot\rho = -3\frac{\dot a}{a}(\rho + p) \; .$$

It shows that the energy density of a component with equation of state $p = w\rho$, $w = \text{const}$ scales as $\rho \propto a^{-3(1+w)}$. As pointed out above, radiation ($w_{\rm rad} = 1/3$) and matter ($w = 0$) scale as $\rho_{\rm rad} \propto a^{-4}$ and $\rho_{\rm M} \propto a^{-3}$, respectively, while the cosmological constant case corresponds to $w_\Lambda = -1$.

Question. Show that for a gas of relativistic particles, $p = \rho/3$.

According to Eqs. (15), (16) and (17), different forms of energy dominate at different cosmological epochs. The present Universe is at the end of the transition from matter domination to Λ domination: the dark energy will 'soon' completely dominate over non-relativistic matter because of the rapid decrease of the energy density of the latter. Conversely, the matter energy density increases as we go backwards in time, and until relatively recently ($z \lesssim 0.3$) it dominated over dark energy density. At even more distant past, the radiation energy density was the highest, as it increases most rapidly backwards in time. The red shift at radiation–matter equality, when the energy densities of radiation and matter were equal, is

$$1 + z_{\rm eq} = \frac{a_0}{a(t_{\rm eq})} = \frac{\Omega_{\rm M}}{\Omega_{\rm rad}} \approx 3500$$

and, using the Friedmann equation, one finds the age of the Universe at equality

$$t_{\rm eq} \approx 50\,000 \text{ years} \; .$$

Note that recombination occurred at matter domination, but rather soon after equality. So, we have the following sequence of the regimes of evolution:

$$\ldots \Longrightarrow \text{Radiation domination} \Longrightarrow \text{Matter domination} \Longrightarrow \Lambda \text{ domination} \; .$$

The dots here denote some cosmological epoch preceding the hot stage of the evolution; as we mentioned in Section 1, we are confident that such an epoch existed, but do not quite know what it was.

2.4 Radiation domination

The epoch of particular interest for our purposes is radiation domination. By inserting $\rho_{\rm rad} \propto a^{-4}$ into the Friedmann equation (11), we obtain

$$\frac{\dot a}{a} = \frac{\text{const}}{a^2} \; .$$

This gives the evolution law

$$a(t) = \text{const} \cdot \sqrt{t} \; . \tag{19}$$

The constant here does not have physical significance, as one can rescale the coordinates $\mathbf{x}$ at some fixed moment of time, thus changing the normalization of a.

There are several points to note regarding the result (19). First, the expansion *decelerates*:

$$\ddot{a} < 0 \ .$$

This property holds also for the matter-dominated epoch, but it does not hold for the domination of the dark energy.

Question. Find the evolution laws, analogous to Eq. (19), for matter- and Λ-dominated Universes. Show that the expansion decelerates, $\ddot{a} < 0$, at matter domination and accelerates, $\ddot{a} > 0$, at Λ domination.

Second, time $t = 0$ is the Big Bang singularity (assuming erroneously that the Universe starts being radiation dominated). The expansion rate

$$H(t) = \frac{1}{2t}$$

diverges as $t \to 0$, and so do the energy density $\rho(t) \propto H^2(t)$ and temperature $T \propto \rho^{1/4}$. Of course, the classical general relativity and usual notions of statistical mechanics (e.g., temperature itself) are not applicable very near the singularity, but our result suggests that in the picture we discuss (hot epoch right after the Big Bang), the Universe starts its classical evolution in a very hot and dense state, and its expansion rate is very high in the beginning. It is customary to consider for illustrational purposes that the relevant quantities in the beginning of the classical expansion take the Planck values, $\rho \sim M_{\rm Pl}^4$, $H \sim M_{\rm Pl}$ etc.

Third, at a given moment of time the size of a causally connected region is finite. Consider signals emitted right after the Big Bang and travelling with the speed of light. These signals travel along the light cone with $\mathrm{d}s = 0$ and hence $a(t)\mathrm{d}x = \mathrm{d}t$. So, the coordinate distance that a signal travels from the Big Bang to time t is

$$x = \int_0^t \frac{\mathrm{d}t}{a(t)} \equiv \eta \ . \tag{20}$$

In the radiation-dominated Universe,

$$\eta = \text{const} \cdot \sqrt{t} \ .$$

The physical distance from the emission point to the position of the signal is

$$l_{\rm H}(t) = a(t)x = a(t) \int_0^t \frac{\mathrm{d}t}{a(t)} = 2t \ .$$

As expected, this physical distance is finite, and it gives the size of a causally connected region at time t. It is called the horizon size (more precisely, the size of the particle horizon). A related property is that an observer at time t can see only the part of the Universe whose current physical size is $l_{\rm H}(t)$. Both at radiation and matter domination one has, modulo a numerical constant of order 1,

$$l_{\rm H}(t) \sim H^{-1}(t) \ . \tag{21}$$

To give an idea of numbers, the horizon size at the present epoch is

$$l_{\rm H}(t_0) \approx 15 \text{ Gpc} \simeq 4.5 \times 10^{28} \text{ cm} \ .$$

Question. Find the proportionality constant in Eq. (21) for a matter-dominated Universe. Is there a particle horizon in a Universe without matter but with positive cosmological constant?

It is convenient to express the Hubble parameter at radiation domination in terms of temperature. The Stefan–Boltzmann law gives for the energy density of a gas of relativistic particles in thermal equilibrium at zero chemical potentials (chemical potentials in the Universe are indeed small)

$$\rho_{\rm rad} = \frac{\pi^2}{30} g_* T^4 \ , \tag{22}$$

with g_* being the effective number of degrees of freedom,

$$g_* = \sum_{\rm bosons} g_i + \frac{7}{8} \sum_{\rm fermions} g_i \ ,$$

where g_i is the number of spin states and the factor $7/8$ is due to Fermi statistics. Hence, the Friedmann equation (11) gives

$$H = \frac{T^2}{M_{\rm Pl}^*} \ , \qquad M_{\rm Pl}^* = \frac{M_{\rm Pl}}{1.66\sqrt{g_*}} \ . \tag{23}$$

One more point has to do with entropy: the cosmological expansion is slow, so that the entropy is conserved (modulo exotic scenarios with large entropy generation). The entropy density in thermal equilibrium is given by

$$s = \frac{2\pi^2}{45} g_* T^3 \ .$$

The conservation of entropy means that the entropy density scales *exactly* as a^{-3},

$$s a^3 = \text{const} \ , \tag{24}$$

while temperature scales *approximately* as a^{-1}. The temperature would scale as a^{-1} if the number of relativistic degrees of freedom would be independent of time. This is not the case, however. Indeed, the value of g_* depends on temperature: at $T \sim 10$ MeV relativistic species are photons, neutrinos, electrons and positrons, while at $T \sim 1$ GeV four flavours of quarks, gluons, muons and τ-leptons are relativistic too. The number of degrees of freedom in the Standard Model at $T \gtrsim 100$ GeV is

$$g_*(100 \text{ GeV}) \approx 100 \ .$$

If there are conserved quantum numbers, such as the baryon number after baryogenesis, their density also scales as a^{-3}. Hence, the time-independent characteristic of, say, the baryon abundance is the baryon-to-entropy ratio

$$\Delta_{\rm B} = \frac{n_{\rm B}}{s} \ .$$

The commonly used baryon-to-photon ratio $\eta_{\rm B}$, Eq. (9), is related to $\Delta_{\rm B}$ by a numerical factor, but this factor depends on time through g_* and stays constant only after $\mathrm{e}^+\mathrm{e}^-$ annihilation, i.e., at $T \lesssim 0.5$ MeV. Numerically,

$$\Delta_{\rm B} = 0.14 \eta_{\rm B,0} = 0.86 \times 10^{-10} \ . \tag{25}$$

3 Dark energy

Before turning to our main topics, let us briefly discuss dark energy. We know very little about this 'substance': our knowledge is summarized in Eqs. (13c) and (18). We also know that dark energy does not clump, unlike dark matter and baryons. It gives rise to the accelerated expansion of the Universe. Indeed, the solution to the Friedmann equation (11) with constant $\rho = \rho_\Lambda$ is

$$a(t) = \mathrm{e}^{H_\Lambda t} \ ,$$

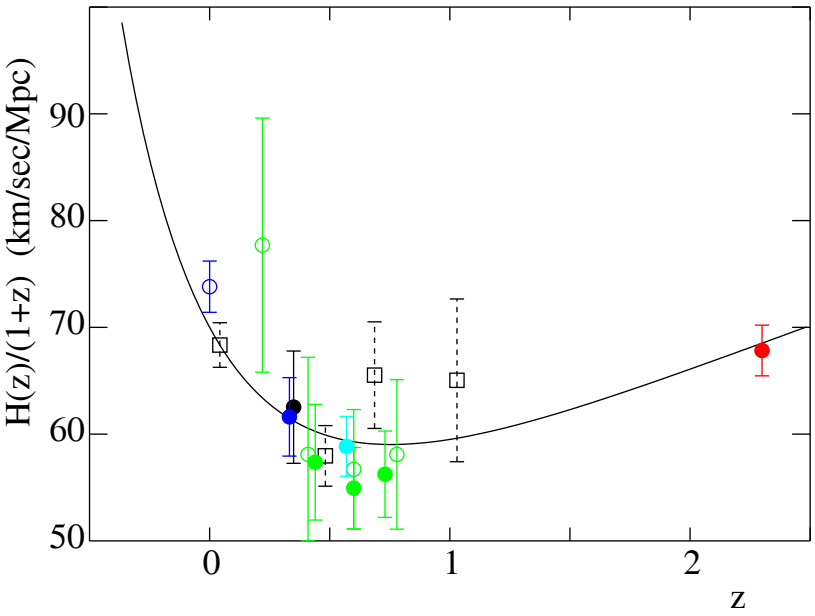


Fig. 5: Observational data on the time derivative of the scale factor as function of red shift z [27]. The change of the behaviour from decreasing to increasing with decreasing z means the change from decelerated to accelerated expansion. The theoretical curve corresponds to a spatially flat Universe with $h = 0.7$ and $\Omega_\Lambda = 0.73$.

where $H_\Lambda = (8\pi\rho_\Lambda/3M_{\rm Pl}^2)^{1/2} = \text{const}$. This gives $\ddot{a} > 0$, unlike at radiation or matter domination. The observational discovery of the accelerated expansion of the Universe was the discovery of dark energy. Recall that early on (substantial z), the Universe was matter dominated, so its expansion was decelerating. The transition from decelerating to accelerating expansion is confirmed by combined observational data, see Fig. 5, which shows the dependence on red shift of the quantity $H(z)/(1+z) = \dot{a}(t)/a_0$.

Question. Find the red shift z at which decelerated expansion turned into an accelerated one.

As a remark, the effective pressure of dark energy or any other component is defined as the (possibly time-dependent) parameter determining the spatial components of the energy–momentum tensor in a locally Lorentz frame ($a = 1$ in the FLRW context),

$$T_{\mu\nu} = \text{diag}\,(\rho, p, p, p)\ .$$

In the case of the cosmological constant, the dark energy density does not depend on time at all:

$$T_{\mu\nu} = \rho_\Lambda \eta_{\mu\nu}\ ,$$

where $\eta_{\mu\nu}$ is the Minkowski tensor. Hence, $w_\Lambda = -1$. One can view this as the characteristic of vacuum, whose energy–momentum tensor must be Lorentz-covariant. As we pointed out above, any deviation from $w = -1$ would mean that we are dealing with something other than vacuum energy density.

The problem with dark energy is that its present value is extremely small by particle-physics standards,

$$\rho_{\rm DE} \approx 4\ \text{GeV m}^{-3} = (2 \times 10^{-3}\ \text{eV})^4\ .$$

In fact, there are two hard problems. One is that particle-physics scales are much larger than the scale relevant to the dark energy density, so the dark energy density is zero to an excellent approximation. Another is that it is non-zero nevertheless, and one has to understand its energy scale. To quantify the first problem, we recall the known scales of particle physics and gravity,

$$\text{Strong interactions}: \qquad \Lambda_{\rm QCD} \sim 1\,\text{GeV}\ ,$$

$$\begin{aligned} \text{Electroweak}: &\qquad M_{\rm W} \sim 100\,\text{GeV}\,, \\ \text{Gravitational}: &\qquad M_{\rm Pl} \sim 10^{19}\,\text{GeV}\,. \end{aligned}$$

Off hand, physics at scale M should contribute to the vacuum energy density as $\rho_\Lambda \sim M^4$, and there is absolutely no reason for vacuum to be as light as it is. The discrepancy here is huge, as one sees from the above numbers.

To elaborate on this point, let us note that the action of gravity plus, say, the Standard Model has the general form

$$S = S_{\rm EH} + S_{\rm SM} - \rho_{\Lambda,0} \int \sqrt{-g}\,\mathrm{d}^4x\,,$$

where $S_{\rm EH} = -(16\pi G)^{-1}\int R\,\sqrt{-g}\,\mathrm{d}^4x$ is the Einstein–Hilbert action of general relativity, $S_{\rm SM}$ is the action of the Standard Model and $\rho_{\Lambda,0}$ is the bare cosmological constant. In order that the vacuum energy density be almost zero, one needs fantastic cancellations between the contributions of the Standard Model fields into the vacuum energy density, on the one hand, and $\rho_{\Lambda,0}$ on the other. For example, we know that quantum chromodynamics (QCD) has a complicated vacuum structure, and one would expect that the energy density of QCD should be of the order of $(1\ \text{GeV})^4$. At least for QCD, one needs a cancellation of the order of 10^{-44}. If one goes further and considers other interactions, the numbers get even worse.

What are the hints from this 'first' cosmological constant problem? There are several options, though not many. One is that the Universe could have a very long prehistory: extremely long. This option has to do with relaxation mechanisms. Suppose that the original vacuum energy density is indeed large, say, comparable to the particle-physics scales. Then there must be a mechanism which can relax this value down to an acceptably small number. It is easy to convince oneself that this relaxation could not happen in the history of the Universe we know of. Instead, the Universe should have a very long prehistory during which this relaxation process might occur. At that prehistoric time, the vacuum in the Universe must have been exactly the same as our vacuum, so the Universe in its prehistory must have been exactly like ours, or almost exactly like ours. Only in that case could a relaxation mechanism work. There are concrete scenarios of this sort [28, 29]. However, at the moment it seems that these scenarios are hardly testable, since this is prehistory.

Another possible hint is towards anthropic selection. The argument that goes back to Weinberg and Linde [30, 31] is that if the cosmological constant were larger, say, by a factor of 100, we simply would not exist: the stars would not have formed because of the fast expansion of the Universe. So, the vacuum energy density may be selected anthropically. The picture is that the Universe may be much, much larger than what we can see, and different large regions of the Universe may have different properties. In particular, vacuum energy density may be different in different regions. Now, we are somewhere in the place where one can live. All the rest is empty of observers, because there the parameters such as vacuum energy density are not suitable for their existence. This is disappointing for a theorist, as this point of view allows for arbitrary tuning of fundamental parameters. It is hard to disprove this option, on the other hand. We do exist, and this is an experimental fact. The anthropic viewpoint may, though hopefully will not, get more support from the LHC, if no or insufficient new physics is found there. Indeed, another candidate for an environmental quantity is the electroweak scale, which is fine tuned in the Standard Model in the same sense as the cosmological constant is fine tuned in gravity (in the Standard Model context, this fine tuning goes under the name of the gauge hierarchy problem).

Turning to the 'second' cosmological constant problem, we note that the scale 10^{-3} eV may be associated with some new light field(s), rather than with vacuum. This implies that ρ_Λ depends on time, i.e., $w_\Lambda \neq -1$ and w_Λ may well depend on time itself. Current data are compatible with time-independent w_Λ equal to -1, but their precision is not particularly high. We conclude that future cosmological observations may shed new light on the field content of the fundamental theory.

4 Dark matter

Unlike dark energy, dark matter experiences the same gravitational force as baryonic matter. It consists presumably of new stable massive particles. These make clumps of mass which constitute most of the mass of galaxies and clusters of galaxies. There are various ways of measuring the contribution of non-baryonic dark matter into the total energy density of the Universe (see Refs. [7–10] for details).

1. The composition of the Universe affects the angular anisotropy and polarization of CMB. Quite accurate CMB measurements available today enable one to measure the total mass density of dark matter.
2. There is direct evidence that dark matter exists in the largest gravitationally bound objects—clusters of galaxies. There are various methods to determine the gravitating mass of a cluster, and even the mass distribution in a cluster, which give consistent results. As an example, the total gravitational field of a cluster, produced by both dark matter and baryons, acts as a gravitational lens for extended light sources behind the cluster. The images of these sources enable one to reconstruct the mass distribution in the cluster. This is shown in Fig. 6. These determinations show that baryons (independently measured through their X-ray emission) make less than $1/3$ of total mass in clusters. The rest is dark matter.

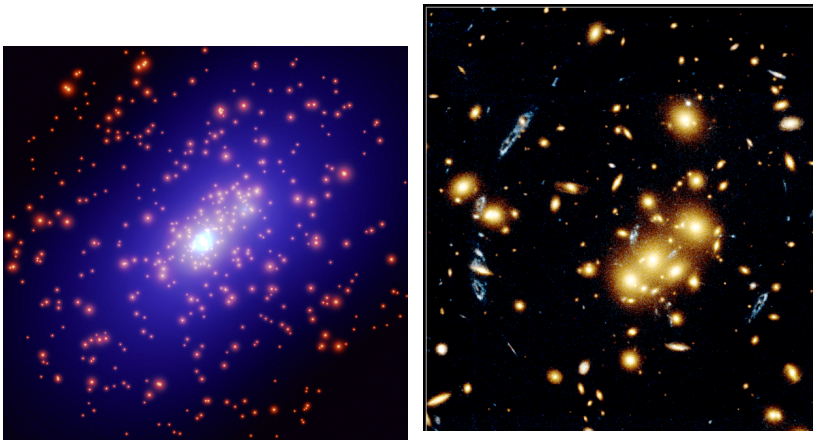


Fig. 6: Cluster of galaxies CL$0024+1654$ [32], acting as gravitational lens. Right-hand panel: cluster in visible light. Round yellow spots are galaxies in the cluster. Elongated blue images are those of one and the same galaxy beyond the cluster. Left-hand panel: reconstructed distribution of gravitating mass in the cluster; brighter regions have larger mass density.

A particularly convincing case is the Bullet Cluster, Fig. 7. Shown are two galaxy clusters that passed through each other. The dark matter and galaxies do not experience friction and thus do not lose their velocities. On the contrary, baryons in hot, X-ray-emitting gas do experience friction and hence get slowed down and lag behind dark matter and galaxies. In this way the baryons (which are mainly in hot gas) and dark matter are separated in space.

3. Dark matter exists also in galaxies. Its distribution is measured by the observations of rotation velocities of distant stars and gas clouds around a galaxy, Fig. 8. Because of the existence of dark matter away from the luminous regions, i.e., in halos, the rotation velocities do not decrease with the distance from the galactic centres; rotation curves are typically flat up to distances exceeding the size of the bright part by a factor of 10 or so. The fact that dark matter halos are so large is explained by the defining property of dark matter particles: they do not lose their energies by emitting photons and, in general, interact with conventional matter very weakly.

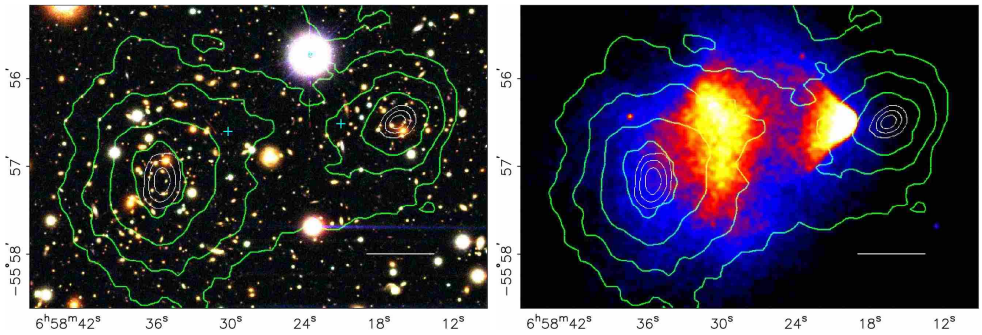


Fig. 7: Observation [33] of the Bullet Cluster 1E0657-558 at $z = 0.296$. Closed lines show the gravitational potential produced mainly by dark matter and measured through gravitational lensing. Bright regions show X-ray emission of hot baryon gas, which makes most of the baryonic matter in the clusters. The length of the white interval is 200 kpc in the comoving frame.

Dark matter is characterized by the mass-to-entropy ratio,

$$\left(\frac{\rho_{\rm DM}}{s}\right)_0 = \frac{\Omega_{\rm DM}\rho_{\rm c}}{s_0} \approx \frac{0.26 \times 5 \times 10^{-6}\ {\rm GeV\ cm^{-3}}}{3000\ {\rm cm^{-3}}} = 4 \times 10^{-10}\ {\rm GeV}\ . \tag{26}$$

This ratio is constant in time since the freeze out of dark matter density: both number density of dark matter particles $n_{\rm DM}$ (and hence their mass density $\rho_{\rm DM} = m_{\rm DM} n_{\rm DM}$) and entropy density get diluted exactly as a^{-3}.

Dark matter is crucial for our existence, for the following reason. Density perturbations in baryon–electron–photon plasma before recombination do not grow because of high pressure, which is mostly due to photons; instead, perturbations are sound waves propagating in plasma with time-independent amplitudes. Hence, in a Universe without dark matter, density perturbations in the baryonic component would start to grow only after baryons decouple from photons, i.e., after recombination. The mechanism

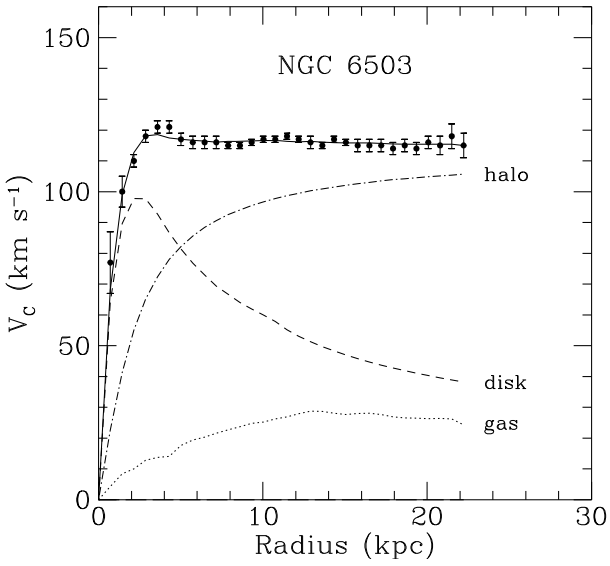


Fig. 8: Rotation velocities of hydrogen gas clouds around the galaxy NGC 6503 [34]. Lines show the contributions of the three main components that produce the gravitational potential. The main contribution at large distances is due to dark matter, labelled 'halo'.

of the growth is pretty simple: an overdense region gravitationally attracts surrounding matter; this matter falls into the overdense region, and the density contrast increases. In the expanding matter-dominated Universe this gravitational instability results in the density contrast growing like $(\delta\rho/\rho)(t) \propto a(t)$. Hence, in a Universe without dark matter, the growth factor for baryon density perturbations would be at most

$$\frac{a(t_0)}{a(t_{\rm rec})} = 1 + z_{\rm rec} = \frac{T_{\rm rec}}{T_0} \approx 10^3 \ . \tag{27}$$

Because of the presence of dark energy, the growth factor is even somewhat smaller. The initial amplitude of density perturbations is very well known from the CMB anisotropy measurements, $(\delta\rho/\rho)_{\rm i} = 5 \times 10^{-5}$. Hence, a Universe without dark matter would still be pretty homogeneous: the density contrast would be in the range of a few per cent. No structure would have been formed, no galaxies, no life. No structure would be formed in future either, as the accelerated expansion due to dark energy will soon terminate the growth of perturbations.

Since dark matter particles decoupled from plasma much earlier than baryons, perturbations in dark matter started to grow much earlier. The corresponding growth factor is larger than (27), so that the dark matter density contrast at galactic and subgalactic scales becomes of order one, perturbations enter the non-linear regime and form dense dark matter clumps at $z = 5$–10. Baryons fall into potential wells formed by dark matter, so dark matter and baryon perturbations develop together soon after recombination. Galaxies get formed in the regions where dark matter was overdense originally. For this picture to hold, dark matter particles must be non-relativistic early enough, as relativistic particles fly through gravitational wells instead of being trapped there. This means, in particular, that neutrinos cannot constitute a considerable part of dark matter.

4.1 Cold and warm dark matter

Currently, the most popular dark matter scenario is cold dark matter, CDM. It consists of particles which get out of *kinetic* equilibrium when they are non-relativistic. For dark matter particles Y which are initially in thermal equilibrium with cosmic plasma, this means that their scattering off other particles switches off at $T = T_{\rm d} \ll m_{\rm Y}$. Since then the dark matter particles move freely, their momenta decrease due to red shift, and they remain non-relativistic until now. Note that the *decoupling* temperature $T_{\rm d}$ may be much lower than the *freeze-out* temperature $T_{\rm f}$ at which the dark matter particles get out of *chemical* equilibrium, i.e., their number in the comoving volume freezes out (because, e.g., their creation and annihilation processes switch off). This is the case for many models with weakly interacting massive particles (WIMPs), a class of dark matter particles we discuss in some detail below. Note also that dark matter particles may never be in thermal equilibrium; this is the case, e.g., for axions.

An alternative to CDM is *warm dark matter*, WDM, whose particles decouple, being relativistic. Let us assume for definiteness that they are in kinetic equilibrium with cosmic plasma at temperature $T_{\rm f}$ when their number density freezes out (thermal relic). After kinetic equilibrium breaks down at temperature $T_{\rm d} \leq T_{\rm f}$, their spatial momenta decrease as a^{-1}, i.e., the momenta are of order T all the time after decoupling. Warm dark matter particles become non-relativistic at $T \sim m$, where m is their mass. Only after that do the WDM perturbations start to grow: as we mentioned above, relativistic particles escape from gravitational potentials, so the gravitational wells get smeared out instead of getting deeper. Before becoming non-relativistic, WDM particles travel the distance of the order of the horizon size; the WDM perturbations therefore are suppressed at those scales. The horizon size at the time $t_{\rm nr}$ when $T \sim m$ is of order

$$l_{\rm H}(t_{\rm nr}) \simeq H^{-1}(T \sim m) = \frac{M^*_{\rm Pl}}{T^2} \sim \frac{M^*_{\rm Pl}}{m^2} \ .$$

Due to the expansion of the Universe, the corresponding length at present is

$$l_0 = l_{\rm H}(t_{\rm nr}) \frac{a_0}{a(t_{\rm nr})} \sim l_{\rm H}(t_{\rm nr}) \frac{T}{T_0} \sim \frac{M_{\rm Pl}}{m T_0} \ , \tag{28}$$

where we neglected (rather weak) dependence on g_*. Hence, in the WDM scenario, structures of sizes smaller than l_0 are less abundant as compared to CDM. Let us point out that l_0 refers to the size of the perturbation in the linear regime; in other words, this is the size of the region from which matter collapses into a compact object.

There is a hint towards the plausibility of warm, rather than cold, dark matter. It is the dwarf-galaxy problem. According to numerical simulations, the CDM scenario tends to overproduce small objects—dwarf galaxies: it predicts hundreds of satellite dwarf galaxies in the vicinity of a large galaxy like the Milky Way, whereas only dozens of satellites have been observed so far. This argument is still controversial, but, if correct, it does suggest that the dark matter perturbations are suppressed at dwarf-galaxy scales. This is naturally the case in the WDM scenario. The present size of a dwarf galaxy is a few kpc, and the density is about 10^6 of the average density in the Universe. Hence, the size l_0 for these objects is of order 100 kpc $\simeq 3 \times 10^{23}$ cm. Requiring that perturbations of this size, but not much larger, are suppressed, we obtain from (28) the estimate for the mass of a dark matter particle

$$\text{WDM} \ : \quad m_{\rm DM} = 3\text{–}10 \ \text{keV} \ . \tag{29}$$

On the other hand, this effect is absent, i.e., dark matter is cold, for

$$\text{CDM} \ : \quad m_{\rm DM} \gg 10 \ \text{keV} \ . \tag{30}$$

Let us recall that these estimates apply to particles that are initially in kinetic equilibrium with cosmic plasma. They do *not* apply in the opposite case; an example is axion dark matter, which is cold despite being of very small axion mass.

4.2 WIMP miracle

There is a simple mechanism of the dark matter generation in the early Universe. It applies to *cold* dark matter. Because of its simplicity and robustness, it is considered by many as a very likely one, and the corresponding dark matter candidates—WIMPs—as the best candidates. Let us describe this mechanism in some detail.

Let us assume that there exists a heavy stable neutral particle Y, and that Y particles can only be destroyed or created via their pair annihilation or creation, with annihilation products being the particles of the Standard Model. The general scenario for the cosmological behaviour of Y particles is as follows. At high temperatures, $T \gg m_{\rm Y}$, the Y particles are in thermal equilibrium with the rest of the cosmic plasma; there are lots of Y particles in the plasma, which are continuously created and annihilate. As the temperature drops below $m_{\rm Y}$, the equilibrium number density decreases. At some 'freeze-out' temperature $T_{\rm f}$, the number density becomes so small that Y particles can no longer meet each other during the Hubble time, and their annihilation terminates. After that the number density of surviving Y particles decreases like a^{-3}, and these relic particles contribute to the mass density in the present Universe.

Let us estimate the properties of Y particles such that they really serve as dark matter. Elementary considerations of mean free path of a particle in gas give for the lifetime of a non-relativistic Y particle in cosmic plasma, $\tau_{\rm ann}$,

$$\langle \sigma_{\rm ann} \cdot v \rangle \cdot \tau_{\rm ann} \cdot n_{\rm Y} \sim 1 \ ,$$

where v is the relative velocity of Y particles, $\sigma_{\rm ann}$ is the annihilation cross-section at velocity v, averaging is over the velocity distribution of Y particles and $n_{\rm Y}$ is the number density. In thermal equilibrium at $T \ll m_{\rm Y}$, the latter is given by the Boltzmann law at zero chemical potential,

$$n_{\rm Y}^{\rm (eq)} = g_{\rm Y} \cdot \left(\frac{m_{\rm Y} T}{2\pi} \right)^{3/2} \mathrm{e}^{-\frac{m_{\rm Y}}{T}} \ , \tag{31}$$

where $g_{\rm Y}$ is the number of spin states of a Y particle. Let us introduce the notation

$$\langle \sigma_{\rm ann} \cdot v \rangle = \sigma_0$$

(in kinetic equilibrium, the left-hand side is the thermal average). If the annihilation occurs in an s-wave, then σ_0 is a constant independent of temperature; for a p-wave it is somewhat suppressed at $T \ll m_{\rm Y}$, namely $\sigma_0 \propto v^2 \propto T/m_{\rm Y}$. A quick way to come to correct estimate is to compare the lifetime with the Hubble time, or the annihilation rate $\Gamma_{\rm ann} \equiv \tau_{\rm ann}^{-1}$ with the expansion rate H. At $T \sim m_{\rm Y}$, the equilibrium density is of order $n_{\rm Y} \sim T^3$, and $\Gamma_{\rm ann} \gg H$ for not too small σ_0. This means that annihilation (and, by reciprocity, creation) of Y pairs is indeed rapid, and Y particles are indeed in complete thermal equilibrium with the plasma. At very low temperature, on the other hand, the equilibrium number density $n_{\rm Y}^{\rm (eq)}$ is exponentially small, and the equilibrium rate is small too, $\Gamma_{\rm ann}^{\rm (eq)} \ll H$. At low temperatures we cannot, of course, make use of the equilibrium formulas: Y particles no longer annihilate (and, by reciprocity, are no longer created), there is no thermal equilibrium with respect to creation–annihilation processes and the number density $n_{\rm Y}$ gets diluted only because of the cosmological expansion.

The freeze-out temperature $T_{\rm f}$ is determined by the relation[1]

$$\tau_{\rm ann}^{-1} \equiv \Gamma_{\rm ann} \simeq H \ , \tag{32}$$

where we use the equilibrium formulas. Making use of the relation (23) between the Hubble parameter and the temperature at radiation domination, we obtain

$$\sigma_0(T_{\rm f}) \cdot n_{\rm Y}(T_{\rm f}) \sim \frac{T_{\rm f}^2}{M_{\rm Pl}^*} \tag{33}$$

or

$$\sigma_0(T_{\rm f}) \cdot g_{\rm Y} \cdot \left(\frac{m_{\rm Y} T_{\rm f}}{2\pi}\right)^{3/2} {\rm e}^{-\frac{m_{\rm Y}}{T_{\rm f}}} \sim \frac{T_{\rm f}^2}{M_{\rm Pl}^*} \ . \tag{34}$$

The latter equation gives the freeze-out temperature, which, up to log–log corrections, is

$$T_{\rm f} \approx \frac{m_{\rm Y}}{\ln(M_{\rm Pl}^* m_{\rm Y} \sigma_0)} \tag{35}$$

(the possible dependence of σ_0 on temperature is irrelevant in the right-hand side: we are doing the calculation in the leading-log approximation anyway). Note that this temperature is somewhat lower than $m_{\rm Y}$ if the relevant microscopic mass scale is much below $M_{\rm Pl}$. This means that Y particles freeze out when they are indeed non-relativistic and get out of kinetic equilibrium at even lower temperature, hence the term 'cold dark matter'. The fact that the annihilation and creation of Y particles terminate at a relatively low temperature has to do with the rather slow expansion of the Universe, which should be compensated for by the smallness of the number density $n_{\rm Y}$.

At the freeze-out temperature, we make use of Eq. (33) and obtain

$$n_{\rm Y}(T_{\rm f}) = \frac{T_{\rm f}^2}{M_{\rm Pl}^* \sigma_0(T_{\rm f})} \ . \tag{36}$$

Note that this density is inversely proportional to the annihilation cross-section (modulo a logarithm). The reason is that for higher annihilation cross-sections, the creation–annihilation processes are longer in equilibrium, and fewer Y particles survive.

[1] In fact, we somewhat oversimplify the analysis here. The chemical equilibrium breaks down slightly earlier than what we find from Eq. (32): the corresponding temperature is obtained by equating the equilibrium creation–annihilation rate $\Gamma_{\rm ann}$ to the rate of evolution of the equilibrium number density (31), rather than to the Hubble parameter H. For $T \ll m_{\rm Y}$, this gives the equation for the temperature

$$\Gamma_{\rm ann} \simeq \frac{\dot{n}_{\rm Y}}{n_{\rm Y}} \simeq -\frac{m_{\rm Y}}{T}\frac{\dot{T}}{T} = \frac{m_{\rm Y}}{T} H(T) \ .$$

This temperature differs by the log–log correction from $T_{\rm f}$ determined from Eq. (34) and, at this temperature, one has $n_{\rm Y} \gg T^2/(M_{\rm Pl}^* \sigma_0)$, cf. Eq. (36). However, below this temperature, the annihilation of Y particles continues, and it terminates at temperature $T_{\rm f}$ determined by Eq. (32), which gives Eqs. (33) and (36). All this gives rise to log–log corrections, which we do not calculate anyway. So, our estimate for the present dark matter mass density remains valid.

Up to a numerical factor of order 1, the number-to-entropy ratio at freeze-out is

$$\frac{n_{\rm Y}}{s} \simeq \frac{1}{g_*(T_{\rm f}) M_{\rm Pl}^* T_{\rm f} \sigma_0(T_{\rm f})} \,. \tag{37}$$

This ratio stays constant until the present time, so the present number density of Y particles is $n_{\rm Y,0} = s_0 \cdot (n_{\rm Y}/s)_{\text{freeze-out}}$, and the mass-to-entropy ratio is

$$\frac{\rho_{\rm Y,0}}{s_0} = \frac{m_{\rm Y} n_{\rm Y,0}}{s_0} \simeq \frac{\ln(M_{\rm Pl}^* m_{\rm Y} \sigma_0)}{g_*(T_{\rm f}) M_{\rm Pl}^* \sigma_0(T_{\rm f})} \simeq \frac{\ln(M_{\rm Pl}^* m_{\rm Y} \sigma_0)}{\sqrt{g_*(T_{\rm f})} M_{\rm Pl} \sigma_0(T_{\rm f})} \,,$$

where we made use of (35). This formula is remarkable. The mass density depends mostly on one parameter, the annihilation cross-section σ_0. The dependence on the mass of a Y particle is through the logarithm and through $g_*(T_{\rm f})$; it is very mild. The value of the logarithm here is between 30 and 40, depending on parameters (this means, in particular, that freeze-out occurs when the temperature drops 30 to 40 times below the mass of a Y particle). Inserting $g_*(T_{\rm f}) \sim 100$, as well as the numerical factor omitted in Eq. (37), and comparing with (26), we obtain the estimate

$$\sigma_0(T_{\rm f}) \equiv \langle \sigma v \rangle (T_{\rm f}) = (1\text{–}2) \times 10^{-36} \text{ cm}^2 \,. \tag{38}$$

This is a weak-scale cross-section, which tells us that the relevant energy scale is TeV. We note in passing that the estimate (38) is quite precise and robust.

If the annihilation occurs in an s-wave, the annihilation cross-section may be parametrized as $\sigma_0 = \alpha^2/M^2$, where α is some coupling constant and M is a mass scale (which may be higher than $m_{\rm Y}$). This parametrization is suggested by the picture of Y-pair annihilation via the exchange by another particle of mass M. With $\alpha \sim 10^{-2}$, the estimate for the mass scale is roughly $M \sim 1$ TeV. Thus, with very mild assumptions, we find that the non-baryonic dark matter may naturally originate from the TeV-scale physics. In fact, what we have found can be understood as an approximate equality between the cosmological parameter, the mass-to-entropy ratio of dark matter and the particle-physics parameters,

$$\text{mass-to-entropy} \simeq \frac{1}{M_{\rm Pl}} \left(\frac{\text{TeV}}{\alpha_{\rm W}} \right)^2 \,.$$

Both are of order 10^{-10} GeV, and it is very tempting to think that this 'WIMP miracle' is not a mere coincidence. If it is not, the dark matter particles should be found at the LHC.

The most prominent candidate for WIMPs is neutralinos of the supersymmetric extensions of the Standard Model. The situation with neutralinos is somewhat tense, however. The point is that the pair annihilation of neutralinos often occurs in the p-wave, rather than the s-wave. This gives the suppression factor in $\sigma_0 \equiv \langle \sigma_{\rm ann} v \rangle$ proportional to $v^2 \sim T_{\rm f}/m_{\rm Y} \sim 1/30$. Hence, neutralinos tend to be overproduced in most of the parameter space of the Minimal Supersymmetric Standard Model (MSSM) and other models. Yet neutralinos remain a good candidate, especially at high $\tan\beta$.

A direct search for dark matter WIMPs is underway in underground laboratories. The idea is that WIMPs orbiting around the centre of our Galaxy with velocity of order 10^{-3} sometimes hit a nucleus in a detector and deposit a small energy in it. These searches have become sensitive to neutralinos, as shown in Fig. 9. Indirect searches for dark matter WIMPs include the search for neutrinos coming from the centres of the Earth and Sun (WIMPs may concentrate and annihilate there), see, e.g., Ref. [36] and positrons and antiprotons in cosmic rays (produced in WIMP annihilations in our Galaxy), see, e.g., Ref. [37]. Collider searches are sensitive to WIMPs too, see Fig. 10. We conclude that the hunt for WIMPs has entered the promising stage.

Question. Estimate the energy deposited in the XENON detector due to elastic scattering of a dark matter WIMP, for WIMP masses 10 GeV, 100 GeV and 1 TeV. Estimate the number of events per kilogram per

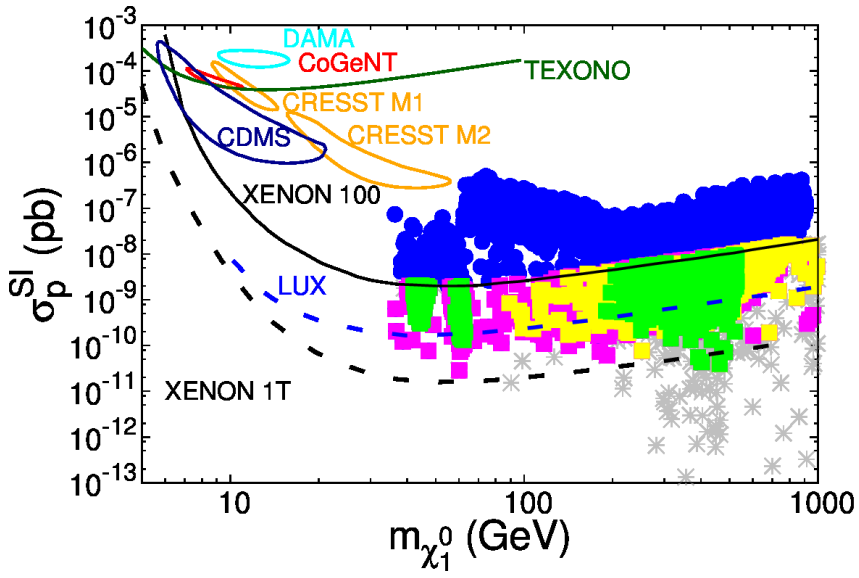


Fig. 9: MSSM predictions for spin-independent elastic neutralino–nucleon cross-section versus neutralino mass and experimentally excluded regions [35]. Shaded regions correspond to MSSM parameters consistent with collider limits and yielding $\Omega_{\rm DM} \approx 0.25$. Regions above the open solid lines are ruled out by direct searches, closed solid curves correspond to regions favoured by experiments indicated. Dashed lines are sensitivities of future direct search experiments LUX and XENON 1T.

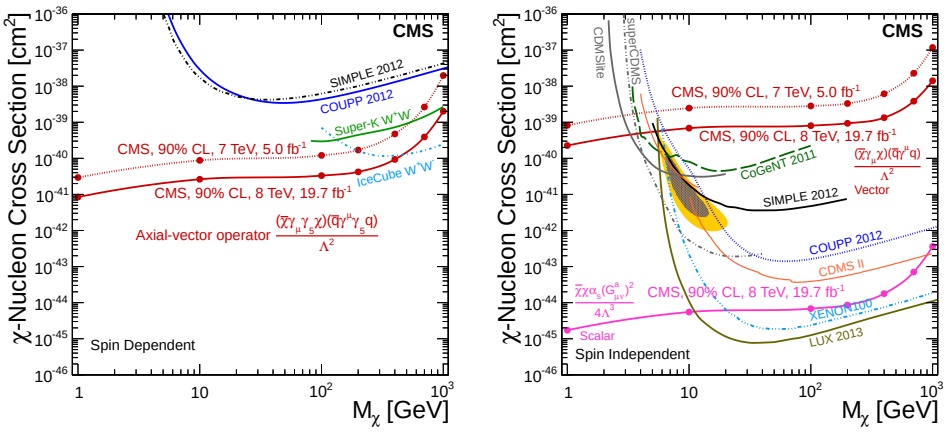


Fig. 10: Excluded regions in the parameter space $(M_{\rm X}, \sigma_{\rm pX})$ [38] for spin-dependent (left) and spin-independent (right) WIMP interactions with nucleons. Regions above the curves are ruled out at 90 % confidence level. CMS denotes searches for WIMPs at the LHC (assuming contact interaction $\mathrm{YYf_1f_2}$, where $\mathrm{f_{1,2}}$ are Standard Model fermions); IceCube and Super-K are searches for neutrinos from WIMP annihilation in the Sun; others are direct searches. The shaded region in the middle of the right-hand panel is favoured by a possible signal at the CDMS experiment.

year for the same masses and elastic cross-sections 10^{-5} pb, 10^{-9} pb and 10^{-8} pb, respectively (see Fig. 9), assuming that the WIMP mass density around the Earth is similar to the average baryon mass density, $\rho_{\rm DM} \sim 0.3$ GeV cm^{-3}, and that $v_{\rm DM} \sim 10^{-3}$.

4.3 Light long-lived particles

Many extensions of the Standard Model contain light scalar or pseudoscalar particles. In some models these new particles are so weakly interacting that their lifetime exceeds the present age of the Universe. Hence, they may serve as dark matter candidates. The best motivated of them is the axion, but there is an entire zoo of axion-like particles.

Let us consider general properties of models with light scalars or pseudoscalars. These particles should interact with the usual matter very weakly, so they must be neutral with respect to the Standard Model gauge interactions. This implies that interactions of scalars S and pseudoscalars P with gauge fields are of the form

$$\mathcal{L}_{SFF} = \frac{C_{SFF}}{4\Lambda} \cdot S F_{\mu\nu} F^{\mu\nu} \ , \quad \mathcal{L}_{PFF} = \frac{C_{PFF}}{8\Lambda} \cdot P F_{\mu\nu} F_{\lambda\rho} \epsilon^{\mu\nu\lambda\rho} \ , \tag{39}$$

where $F_{\mu\nu}$ is the field strength of the $\mathrm{SU}(3)_\mathrm{c}$, $\mathrm{SU}(2)_\mathrm{W}$ or $\mathrm{U}(1)_\mathrm{Y}$ gauge group. The parameter Λ has dimension of mass and can be interpreted as the scale of new physics related to an S and/or a P particle. This parameter has to be large; then the interactions of S and P with gauge bosons are indeed weak at low energies. Because of that, the Lagrangians (39) contain gauge-invariant operators of the lowest possible dimension. Dimensionless constants C_{SFF} and C_{PFF} are typically numbers of order 1. The terms (39) describe interactions of (pseudo)scalars with pairs of photons, gluons as well as with $\mathrm{Z}\gamma$, ZZ and $\mathrm{W}^+\mathrm{W}^-$ pairs.

Interactions with fermions can also be written on symmetry grounds. Since S and P are singlets under $\mathrm{SU}(3)_\mathrm{c} \times \mathrm{SU}(2)_\mathrm{W} \times \mathrm{U}(1)_\mathrm{Y}$, no combinations like $S\bar{f}f$ or $P\bar{f}\gamma^5 f$ are gauge invariant, so they cannot appear in the Lagrangian (hereafter f denotes the Standard Model fermions). Gauge-invariant operators of the lowest dimension have the form $H\bar{f}f$, where H is the Englert–Brout–Higgs field. Hence, the interactions with fermions are

$$\mathcal{L}SHff = \frac{Y_{SHff}}{\Lambda} \cdot SH\bar{f}f \ , \quad \mathcal{L}PHff = \frac{Y_{PHff}}{\Lambda} \cdot PH\bar{f}\gamma^5 f \ .$$

It often happens that the couplings Y_{SHff} and Y_{SPff} are of the order of the Standard Model Yukawa couplings, so upon electroweak symmetry breaking the low-energy Lagrangians have the following structure:

$$\mathcal{L}Sff = \frac{C_{Sff} m_f}{\Lambda} \cdot S\bar{f}f \ , \quad \mathcal{L}Pff = \frac{C_{Pff} m_f}{\Lambda} \cdot P\bar{f}\gamma^5 f \ , \tag{40}$$

where we assume that the dimensionless couplings C_{Sff} and C_{Pff} are also of order 1.

Making use of Eqs. (39) and (40), we estimate the partial widths of decays of P and S into the Standard Model particles:

$$\Gamma_{P(S)\to AA} \sim \frac{m_{P(S)}^3}{64\pi\Lambda^2} \ , \quad \Gamma_{P(S)\to ff} \sim \frac{m_f^2 m_{P(S)}}{8\pi\Lambda^2} \ , \tag{41}$$

where A denotes vector bosons. By requiring that the lifetime of the new particles exceeds the present age of the Universe, $\tau_{S(P)} = \Gamma_{S(P)}^{-1} > H_0^{-1}$, we find a bound on the mass of the dark matter candidates,

$$m_{P(S)} < \left(16\pi\Lambda^2 H_0\right)^{1/3} \ . \tag{42}$$

Assuming that the new physics scale is below the Planck scale, $\Lambda < M_\mathrm{Pl}$, we obtain an (almost) model-independent bound,

$$m_{P(S)} < 100 \text{ MeV} \ . \tag{43}$$

Hence, the kinematically allowed decays are $P(S) \to \gamma\gamma$, $P(S) \to \nu\bar{\nu}$ and $P(S) \to \mathrm{e}^+\mathrm{e}^-$. It follows from Eq. (41) that the two-photon decay mode dominates, unless the mass of the new particle is close to that of the electron.

Let us now consider generation of relic (pseudo)scalars in the early Universe. There are several generation mechanisms; one of them is fairly generic for the class of models we discuss. This is generation in decays of condensates (we will consider another mechanism later, in the model with axions). The picture is as follows. Let some scalar field ϕ be in a condensate in the early Universe. The condensate can be viewed as a collection of ϕ particles at rest. Equivalently, the condensate is the homogeneous scalar field that oscillates at relatively late times, when $m_\phi > H$. Let both particles, ϕ and S, interact with matter so weakly that they never get into thermal equilibrium, and let the interaction between ϕ and S have the form $\mu\phi S^2/2$, where μ is the coupling constant. Then the width of the decay $\phi \to SS$ is estimated as

$$\Gamma_{\phi\to SS} \sim \frac{\mu^2}{16\pi m_\phi} \,. \tag{44}$$

If the widths of other decay channels do not exceed the value (44), the decay of the ϕ condensate occurs at a temperature T_ϕ determined by

$$\Gamma_{\phi\to SS} \sim H(T_\phi) = \frac{T_\phi^2}{M_{\rm Pl}^*} \,.$$

Let the energy density of the ϕ condensate at that time be equal to ρ_ϕ, so that the number density of decaying ϕ particles is $n_\phi \sim \rho_\phi/m_\phi$. Immediately after the epoch of ϕ-particle decays, the number density of S particles is of order $\epsilon\rho_\phi/m_\phi$, where ϵ is the fraction of the condensate that decayed into S particles. After S particles become non-relativistic, their mass density is of order

$$\rho_S \sim \epsilon\rho_\phi \cdot \frac{m_S T^3}{m_\phi T_\phi^3} \,,$$

where we omitted the dependence on g_* for simplicity. In this way we estimate the mass fraction of S particles today,

$$\Omega_S = \frac{\rho_S}{\rho_{\rm c}} \sim \frac{m_S T_0^3}{\rho_{\rm c}} \cdot \frac{\epsilon\rho_\phi}{m_\phi T_\phi^3} \sim 0.2 \cdot \left(\frac{m_{\rm S}}{1\ {\rm eV}}\right) \cdot \frac{\epsilon\rho_\phi}{m_\phi T_\phi^3} \,. \tag{45}$$

With an appropriate choice of parameters, the correct value $\Omega_S \simeq 0.2$ can indeed be obtained. We note that the last factor on the right-hand side of Eq. (45) must be small.

4.4 Axions

Let us now turn to a concrete class of models with Peccei–Quinn symmetry and axions. This symmetry provides a solution to the *strong CP-problem,* and the existence of axions is an inevitable consequence of the construction.

The strong CP-problem [39–41] emerges in the following way. One can extend the Standard Model Lagrangian by adding the following term:

$$\Delta L = \frac{\alpha_{\rm s}}{8\pi} \cdot \theta_0 \cdot G^{\rm a}_{\mu\nu}\tilde{G}^{\mu\nu\ {\rm a}} \,, \tag{46}$$

where $\alpha_{\rm s}$ is the SU(3)$_{\rm c}$ gauge coupling, $G^{\rm a}_{\mu\nu}$ is the gluon field strength, $\tilde{G}^{\mu\nu\ {\rm a}} = \frac{1}{2}\epsilon^{\mu\nu\lambda\rho}G^{\rm a}_{\lambda\rho}$ is the dual tensor and θ_0 is an arbitrary dimensionless parameter (the factor $\alpha_{\rm s}/(8\pi)$ is introduced for later convenience). The interaction term (46) is invariant under gauge symmetries of the Standard Model, but it violates P and CP. The term (46) is a total derivative, so it does not contribute to the classical field equations, and its contribution to the action is reduced to the surface integral. For any perturbative gauge field configurations (small perturbations about $G^{\rm a}_\mu = 0$), this contribution is equal to zero. However, this is not the case for configurations of instanton type. This means that CP is violated in QCD at the non-perturbative level.

Furthermore, quantum effects due to quarks give rise to the anomalous term in the Lagrangian, which has the same form as Eq. (46) with proportionality coefficient determined by the phase of the quark mass matrix $\hat{M}_{\rm q}$. The latter enters the Lagrangian as

$$\mathcal{L}{\rm m} = \bar{q}_{\rm L}\hat{M}_{\rm q}q_{\rm R} + {\rm h.c.}$$

By chiral rotation of quark fields, one makes quark masses real (i.e., physical), but that rotation induces a new term in the Lagrangian,

$$\Delta\mathcal{L}{\rm m} = \frac{\alpha_{\rm s}}{8\pi}\cdot{\rm Arg}\left({\rm Det}\hat{M}_{\rm q}\right)\cdot G^{\rm a}_{\mu\nu}\tilde{G}^{\mu\nu\ {\rm a}}\ . \tag{47}$$

There is no reason to think that ${\rm Arg}\left({\rm Det}\hat{M}_{\rm q}\right) = 0$. Neither there is a reason to think that the 'tree-level' term (46) and the anomalous contribution (47) cancel each other. Indeed, the former term is there even in the absence of quarks, while the latter comes from the Yukawa sector, as the quark masses are due to their Yukawa interactions with the Englert–Brout–Higgs field.

Thus, the Standard Model Lagrangian should contain the term

$$\Delta\mathcal{L}\theta = \frac{\alpha_{\rm s}}{8\pi}\left(\theta_0 + {\rm Arg}\left({\rm Det}\hat{M}_{\rm q}\right)\right)G^{\rm a}_{\mu\nu}\tilde{G}^{\mu\nu\ {\rm a}} \equiv \frac{\alpha_{\rm s}}{8\pi}\cdot\theta\cdot G^{\rm a}_{\mu\nu}\tilde{G}^{\mu\nu\ {\rm a}}\ . \tag{48}$$

This term violates CP, and off hand the parameter θ is of order 1.

The term (48) has non-trivial phenomenological consequences. One is that it generates the electric dipole moment (EDM) of the neutron, $d_{\rm n}$, which is estimated as [42]

$$d_{\rm n} \sim \theta\times 10^{-16}\,e\ {\rm cm}\ . \tag{49}$$

The neutron EDM has not been found experimentally, and the searches place a strong bound

$$d_{\rm n} \lesssim 3\times 10^{-26}\,e\ {\rm cm}\ . \tag{50}$$

This leads to the bound on the parameter θ,

$$|\theta| < 0.3\times 10^{-9}\ .$$

The problem to explain such a small value of θ is precisely the strong CP-problem.

A solution to this problem does not exist within the Standard Model. The solution is offered by models with axions. These models make use of the following observation. If at the classical level the quark Lagrangian is invariant under axial symmetry ${\rm U}(1)_A$ such that

$$q_{\rm L}\to {\rm e}^{{\rm i}\beta}q_{\rm L}\ ,\qquad q_{\rm R}\to {\rm e}^{-{\rm i}\beta}q_{\rm R}\ , \tag{51}$$

then the θ term would be rotated away by applying this transformation. This global symmetry is called the Peccei–Quinn (PQ) symmetry [43], ${\rm U}(1)_{\rm PQ}$. There is no PQ symmetry in the Standard Model, but one can extend the Standard Model in such a way that the classical Lagrangian is invariant under the PQ symmetry. Quark masses are not invariant under the PQ transformations (51), so PQ symmetry is *spontaneously broken*. At the classical level, this leads to the existence of a massless Nambu–Goldstone field $a(x)$, an axion. As for any Nambu–Goldstone field, its properties are determined by its transformation law under the PQ symmetry:

$$a(x)\to a(x) + \beta\cdot f_{\rm PQ}\ , \tag{52}$$

where β is the same parameter as in Eq. (51) and $f_{\rm PQ}$ is a constant of dimension of mass, the energy scale of ${\rm U}(1)_{\rm PQ}$ symmetry breaking. The mass terms in the low-energy quark Lagrangian must be

symmetric under the transformations (51) and (52), so the quark and axion fields enter the Lagrangian in the combination

$$\mathcal{L}\mathrm{m} = \bar{q}_\mathrm{R} m_\mathrm{q} \mathrm{e}^{-2\mathrm{i}\frac{a}{f_\mathrm{PQ}}} q_\mathrm{L} + \mathrm{h.c.} \tag{53}$$

Making use of Eq. (47), we find that at the quantum level the low-energy Lagrangian contains the term

$$\mathcal{L}\mathrm{a} = C_\mathrm{g} \frac{\alpha_\mathrm{s}}{8\pi} \cdot \frac{a}{f_\mathrm{PQ}} G^\mathrm{a}_{\mu\nu} \tilde{G}^{\mu\nu\ \mathrm{a}} \ , \tag{54}$$

where the constant C_g is of order 1; it is determined by PQ charges of quarks. Clearly, PQ symmetry (51) and (52) is *explicitly* broken by quantum effects of QCD, and an axion is a *pseudo*-Nambu–Goldstone boson.

Hence, the θ parameter multiplying the operator $G^\mathrm{a}_{\mu\nu}\tilde{G}^{\mu\nu\ \mathrm{a}}$ obtains a shift depending on the space–time point and proportional to the axion field,

$$\theta \to \bar{\theta}(x) = \theta + C_\mathrm{g} \frac{a(x)}{f_\mathrm{PQ}} \ . \tag{55}$$

Strong interactions would conserve CP provided the axion vacuum expectation value is such that $\langle\bar{\theta}\rangle = 0$. The QCD effects indeed do the job. They generate a non-vanishing quark condensate $\langle\bar{q}q\rangle \sim \Lambda^3_\mathrm{QCD}$ at the QCD energy scale $\Lambda_\mathrm{QCD} \sim 200$ MeV. This condensate breaks chiral symmetry and in turn generates the axion effective potential

$$V_\mathrm{a} \sim -\frac{1}{2}\bar{\theta}^2 \frac{m_\mathrm{u} m_\mathrm{d}}{m_\mathrm{u} + m_\mathrm{d}} \langle\bar{q}q\rangle + \mathcal{O}(\bar{\theta}^4) \simeq \frac{1}{8}\bar{\theta}^2 \cdot m_\pi^2 f_\pi^2 + \mathcal{O}(\bar{\theta}^4) \ , \tag{56}$$

where $m_\pi = 135$ MeV and $f_\pi = 93$ MeV are pion mass and decay constant. In fact, the axion potential must be periodic in θ with period 2π, so the expression (56) is valid for small θ only. The potential has the minimum at $\langle\bar{\theta}\rangle = 0$, so the strong CP-problem finds an elegant solution. It follows from Eqs. (55) and (56) that the axion has a mass

$$m_\mathrm{a} \approx C_\mathrm{g} \frac{m_\pi f_\pi}{2 f_\mathrm{PQ}} \ , \tag{57}$$

i.e., it is indeed a *pseudo*-Nambu–Goldstone boson.

There are various ways to implement the PQ mechanism. One is to introduce two Englert–Brout–Higgs doublets and choose the Yukawa interaction as

$$Y^\mathrm{d} \bar{Q}_\mathrm{L} H_1 D_\mathrm{R} + Y^\mathrm{u} \bar{Q}_\mathrm{L} \mathrm{i}\tau^2 H_2^* U_\mathrm{R} \ . \tag{58}$$

The two scalar fields transform under the $\mathrm{U}(1)_\mathrm{PQ}$ transformation (51) as follows:

$$H_1 \to \mathrm{e}^{2\mathrm{i}\beta} H_1 \ , \quad H_2 \to \mathrm{e}^{-2\mathrm{i}\beta} H_2 \ .$$

This ensures $\mathrm{U}(1)_\mathrm{PQ}$ invariance of the Lagrangian (58) and hence the absence of the θ term. Both scalars acquire vacuum expectation values v_1 and v_2. If no other new fields are added, we arrive at the Weinberg–Wilczek model [44, 45]. In that case, the axion field θ is the relative phase of H_1 and H_2, and the PQ scale equals the electroweak scale:

$$f_\mathrm{PQ} = 2\sqrt{v_1^2 + v_2^2} = 2v_\mathrm{SM} = 2 \times 246 \text{ GeV} \ .$$

The axion is quite heavy, $m_\mathrm{a} \sim 15$ keV, and its interaction with quarks, gluons and photons is too strong. Because of that, the Weinberg–Wilczek axion is experimentally ruled out.

This problem is solved in the Dine–Fischler–Srednicki–Zhitnitsky (DFSZ) model [46, 47] by adding a complex scalar field S which is a singlet under the Standard Model gauge group. Its interactions involve PQ invariants

$$S^\dagger S \ , \quad H_1^\dagger H_2 \cdot S^2 \ .$$

The field S transforms under $\mathrm{U}(1)_{\mathrm{PQ}}$ as $S \to \mathrm{e}^{2\mathrm{i}\beta} S$. The axion field is now a linear combination of the phases of fields H_1, H_2 and S and

$$f_{\mathrm{PQ}} = 2\sqrt{v_1^2 + v_2^2 + v_{\mathrm{s}}^2}\,, \tag{59}$$

where v_{s} is the vacuum expectation value of the field S. The latter can be large, so it is clear from Eq. (59) that the mass of the axion is small and, most importantly, its couplings to the Standard Model fields are weak: these couplings are inversely proportional to $f_{\mathrm{PQ}} \sim v_{\mathrm{s}}$. The DFSZ axion interacts with both quarks and leptons.

Another approach is called the Kim–Shifman–Vainshtein–Zakharov (KSVZ) mechanism [48, 49]. It does not require more than one Englert–Brout–Higgs field of the Standard Model. The mechanism makes use of additional quark fields Ψ_{R} and Ψ_{L}, which are triplets under $\mathrm{SU}(3)_{\mathrm{c}}$ and singlets under $\mathrm{SU}(2)_{\mathrm{W}} \times \mathrm{U}(1)_{\mathrm{Y}}$. Only these quarks transform non-trivially under $\mathrm{U}(1)_{\mathrm{PQ}}$, while the usual quarks have zero PQ charge. One also introduces a complex scalar field S, which is a singlet under the Standard Model gauge group. One writes the PQ-invariant Yukawa interaction of the new fields,

$$L = y_\Psi S \bar{\Psi}_{\mathrm{R}} \Psi_{\mathrm{L}} + \text{h.c.}\,,$$

so that S again transforms under $\mathrm{U}(1)_{\mathrm{PQ}}$ as $S \to \mathrm{e}^{2\mathrm{i}\beta} S$. PQ symmetry is spontaneously broken by the vacuum expectation value $\langle S \rangle = v_{\mathrm{s}}/\sqrt{2}$. The axion here is the phase of the field S; therefore,

$$f_{\mathrm{PQ}} = 2v_{\mathrm{s}}\,. \tag{60}$$

The KSVZ model does not contain an explicit interaction of an axion with the usual quarks and leptons.

To summarize, an axion is a light particle whose interactions with the Standard Model fields are very weak. The latter property relates to the fact that it is a pseudo-Nambu–Goldstone boson of a global symmetry spontaneously broken at the high-energy scale $f_{\mathrm{PQ}} \gg M_{\mathrm{W}}$. As for any Nambu–Goldstone field, the interactions of an axion with quarks and leptons are described by the generalized Goldberger–Treiman formula

$$\mathcal{L}\mathrm{af} = \frac{1}{f_{\mathrm{PQ}}} \cdot \partial_\mu a \cdot J^\mu_{\mathrm{PQ}}\,. \tag{61}$$

Here

$$J^\mu_{\mathrm{PQ}} = \sum_{\mathrm{f}} e_{\mathrm{f}}^{(\mathrm{PQ})} \cdot \bar{f}\gamma^\mu\gamma^5 f\,. \tag{62}$$

The contributions of fermions to the current J^μ_{PQ} are proportional to their PQ charges $e_{\mathrm{f}}^{(\mathrm{PQ})}$; these charges are model-dependent. In accord with Eq. (53), the action (61) can be integrated by parts and we obtain instead

$$\begin{aligned}\mathcal{L}\mathrm{af} &= -\frac{1}{f_{\mathrm{PQ}}} \cdot a \cdot \partial_\mu J^\mu_{\mathrm{PQ}} \\ &= -\frac{a}{f_{\mathrm{PQ}}} \cdot \sum_f 2e_f^{(\mathrm{PQ})} m_f \cdot \bar{f}\gamma^5 f\,.\end{aligned} \tag{63}$$

Besides the interaction (61), there are also interactions of axions with gluons, see Eq. (54), and photons,

$$\mathcal{L}\mathrm{ag} = C_{\mathrm{g}} \frac{\alpha_{\mathrm{s}}}{8\pi} \cdot \frac{a}{f_{\mathrm{PQ}}} \cdot G^{\mathrm{a}}_{\mu\nu} \tilde{G}^{\mu\nu\,\mathrm{a}}\,, \qquad \mathcal{L}\mathrm{a}\gamma = C_\gamma \frac{\alpha}{8\pi} \cdot \frac{a}{f_{\mathrm{PQ}}} \cdot F_{\mu\nu} \tilde{F}^{\mu\nu}\,, \tag{64}$$

where the dimensionless constants C_{g} and C_γ are also model-dependent and, generally speaking, are of order 1. The interaction terms (63) and (64) indeed have the form (39) and (40), i.e., models with axions

belong to the class of models with light, weakly interacting pseudoscalars. The axion mass, however, is not a free parameter: we find from Eq. (57) that

$$m_{\rm a} \approx m_\pi \cdot \frac{f_\pi}{2f_{\rm PQ}} \approx 0.6 \text{ eV} \cdot \left(\frac{10^7 \text{ GeV}}{f_{\rm PQ}}\right) . \tag{65}$$

The main decay channel of the light axion is decay into two photons. The lifetime $\tau_{\rm a}$ is found from Eq. (41) by setting $\Lambda = 2\pi f_{\rm PQ}/\alpha$ and using Eq. (65),

$$\tau_{\rm a} = \frac{1}{\Gamma_{{\rm a}\to\gamma\gamma}} = \frac{64\pi^3 m_\pi^2 f_\pi^2}{\alpha^2 m_{\rm a}^5} \simeq 4 \times 10^{24} \text{ s} \cdot \left(\frac{\text{eV}}{m_{\rm a}}\right)^5 .$$

By requiring that this lifetime exceeds the age of the Universe, $\tau_{\rm a} > t_0 \approx 14$ billion years, we find the bound on the mass of the axion as a dark matter candidate,

$$m_{\rm a} < 25 \text{ eV} . \tag{66}$$

There are astrophysical bounds on the strength of axion interactions $f_{\rm PQ}^{-1}$ and hence on the axion mass. Axions in theories with $f_{\rm PQ} \lesssim 10^9$ GeV, which are heavier than 10^{-2} eV, would be intensely produced in stars and supernovae explosions. This would lead to contradictions with observations. So, we are left with very light axions, $m_{\rm a} \lesssim 10^{-2}$ eV.

As far as dark matter is concerned, thermal production of axions is irrelevant. There are at least two mechanisms of axion production in the early Universe that can provide not only right axion abundance but also small initial velocities of axions. The latter property makes an axion a *cold* dark matter candidate, despite its very small mass. One mechanism has to do with decays of global strings [50]—topological defects that exist in theories with spontaneously broken global U(1) symmetry ($\rm U(1)_{PQ}$ in our case; for a discussion of this mechanism, see, e.g., Ref. [51]). Another mechanism employs an axion condensate [52–54], an homogeneous axion field that oscillates in time after the QCD epoch. This is called the axion misalignment mechanism. Let us consider the second mechanism in some detail.

As we have seen in Eq. (56), the axion potential is proportional to the quark condensate $\langle \bar{q}q \rangle$. This condensate breaks chiral symmetry. The chiral symmetry is in fact restored at high temperatures Hence, one expects that the axion potential is negligibly small at $T \gg \Lambda_{\rm QCD}$. This is indeed the case: the effective potential for the field $\bar\theta = \theta + a/f_{\rm PQ}$ vanishes at high temperatures, and this field can take any value,

$$\bar\theta_{\rm i} \in [0\,, 2\pi) \ ,$$

where we recall that the field $\bar\theta$ is a phase. There is no reason to think that the initial value $\bar\theta_{\rm i}$ is zero. As the temperature decreases, the axion mass $m(T)$ starts to get generated, so that

$$m_{\rm a}(T) \simeq 0 \qquad \text{at } T \gg \Lambda_{\rm QCD} ,$$
$$m_{\rm a}(T) \simeq m_{\rm a} \qquad \text{at } T \ll \Lambda_{\rm QCD} .$$

Hereafter $m_{\rm a}$ denotes the zero-temperature axion mass. As the mass increases, at some point the field $\bar\theta$, remaining homogeneous, starts to roll down from $\bar\theta_{\rm i}$ towards its value $\bar\theta = 0$ at the minimum of the potential. The axion field practically does not evolve when $m_{\rm a}(T) \ll H(T)$ and at the time when $m_{\rm a}(T) \sim H(T)$ it starts to oscillate. Let us estimate the present energy density of the axion field in this picture, without using the concrete form of the function $m(T)$.

The oscillations start at the time $t_{\rm osc}$ when

$$m_{\rm a}(t_{\rm osc}) \sim H(t_{\rm osc}). \tag{67}$$

At this time, the energy density of the axion field is estimated as

$$\rho_{\rm a}(t_{\rm osc}) \sim m_{\rm a}^2(t_{\rm osc}) f_{\rm PQ}^2 \bar\theta_{\rm i}^2 .$$

The oscillating axion field is the same thing as a collection of axions at rest. Their number density at the beginning of oscillations is estimated as

$$n_{\rm a}(t_{\rm osc}) \sim \frac{\rho_{\rm a}(t_{\rm osc})}{m_{\rm a}(t_{\rm osc})} \sim m_{\rm a}(t_{\rm osc}) f_{\rm PQ}^2 \bar{\theta}_{\rm i}^2 \sim H(t_{\rm osc}) f_{\rm PQ}^2 \bar{\theta}_{\rm i}^2 \ .$$

This number density, as any number density of non-relativistic particles, then decreases as a^{-3}.

The axion-to-entropy ratio at time $t_{\rm osc}$ is

$$\frac{n_{\rm a}}{s} \sim \frac{H(t_{\rm osc}) f_{\rm PQ}^2}{\frac{2\pi^2}{45} g_* T_{\rm osc}^3} \cdot \bar{\theta}_{\rm i}^2 \simeq \frac{f_{\rm PQ}^2}{\sqrt{g_*} T_{\rm osc} M_{\rm Pl}} \cdot \bar{\theta}_{\rm i}^2 \ ,$$

where we use the usual relation $H = 1.66\sqrt{g_*}T^2/M_{\rm Pl}$. The axion-to-entropy ratio remains constant after the beginning of oscillations, so the present mass density of axions is

$$\rho_{\rm a,0} = \frac{n_{\rm a}}{s} m_{\rm a} s_0 \simeq \frac{m_{\rm a} f_{\rm PQ}^2}{\sqrt{g_*} T_{\rm osc} M_{\rm Pl}} s_0 \cdot \bar{\theta}_{\rm i}^2 \ . \tag{68}$$

In fact, it is a decreasing function of $m_{\rm a}$. Indeed, $f_{\rm PQ}$ is inversely proportional to $m_{\rm a}$, see Eq. (57); at the same time, the axion obtains its mass near the epoch of QCD transition, i.e., at $T \sim \Lambda_{\rm QCD}$, so $T_{\rm osc}$ depends on $m_{\rm a}$ rather weakly.

To obtain a simple estimate, let us set $T_{\rm osc} \sim \Lambda_{\rm QCD} \simeq 200$ MeV and make use of Eq. (57) with $C_{\rm g} \sim 1$. We find

$$\Omega_{\rm a} \equiv \frac{\rho_{\rm a,0}}{\rho_{\rm c}} \simeq \left(\frac{10^{-6}\ {\rm eV}}{m_{\rm a}} \right) \bar{\theta}_{\rm i}^2 \ . \tag{69}$$

The natural assumption about the initial phase is $\bar{\theta}_{\rm i} \sim \pi/2$. Hence, an axion of mass 10^{-5}–10^{-6} eV is a good dark matter candidate. Note that an axion of lower mass $m_{\rm a} < 10^{-6}$ eV may also serve as a dark matter particle, if for some reason the initial phase $\bar{\theta}_{\rm i}$ is much smaller than $\pi/2$. This is *cold* dark matter: the oscillating field corresponds to axions at rest.

A more precise estimate is obtained by taking into account the fact that that the axion mass smoothly depends on temperature:

$$\Omega_{\rm a} \simeq 0.2 \cdot \bar{\theta}_{\rm i}^2 \cdot \left(\frac{4 \times 10^{-6}\ {\rm eV}}{m_{\rm a}} \right)^{1.2} \ .$$

We see that our crude estimate (69) is fairly accurate. Interestingly, the string mechanism of the axion production leads to the same parametric dependence of $\Omega_{\rm a}$ on the axion mass.

Search for dark matter axions with mass $m_{\rm a} \sim 10^{-5}$–10^{-6} eV is difficult, but not impossible. One way is to search for axion–photon conversion in a resonator cavity filled with a strong magnetic field. Indeed, in the background magnetic field the axion–photon interaction (second term in Eq. (64)) leads to the conversion a $\to \gamma$, and the axions of mass 10^{-5}–10^{-6} eV are converted to photons of frequency $m/(2\pi) = 2$–0.2 GHz (radio waves). Bounds on the dark matter axions are shown in Fig. 11.

4.5 Warm dark matter: sterile neutrinos and light gravitinos

As we discussed in Section 4.1, there are arguments, albeit not particularly strong, that favour warm, rather than cold, dark matter. If WDM particles are thermal relics, i.e., if they were in kinetic equilibrium at some epoch in the early Universe, then their mass should be in the range 3–10 keV. Reasonably well motivated particles of this mass are sterile neutrinos and gravitinos.

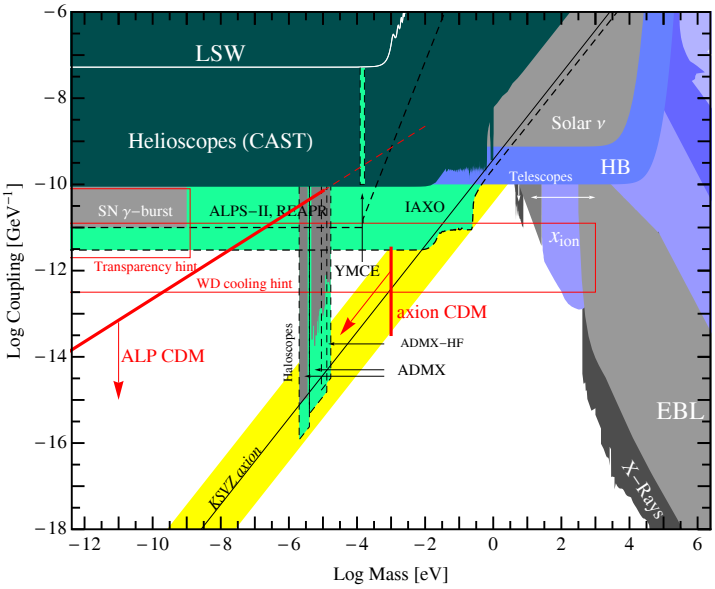


Fig. 11: Bounds on dark matter axions: axion–photon coupling versus axion mass [55]. Inclined straight line labelled 'KSVZ axion' is the prediction of the KSVZ model, shaded region along this line is the range of predictions of other axion models. Region below the line labelled ALP CDM is the range of predictions of other reasonably motivated models with axion-like particles as dark matter candidates. Dashed lines show the sensitivities of future experiments.

4.5.1 Sterile neutrinos

Sterile neutrinos are most probably required for giving masses to ordinary, 'active' neutrinos. The masses of sterile neutrinos cannot be predicted theoretically. Although sterile neutrinos of WDM mass $m_{\nu_{\rm s}} = 3$–10 keV are not particularly plausible from the particle-physics prospective, they are not pathological either. In the simplest models the creation of sterile neutrino states $|\nu_{\rm s}\rangle$ in the early Universe occurs due to their mixing with active neutrinos $|\nu_\alpha\rangle$, $\alpha = {\rm e}, \mu, \tau$. In the approximation of mixing between two states only, we have

$$|\nu_\alpha\rangle = \cos\theta|\nu_1\rangle + \sin\theta|\nu_2\rangle \ , \qquad |\nu_{\rm s}\rangle = -\sin\theta|\nu_1\rangle + \cos\theta|\nu_2\rangle \ ,$$

where $|\nu_\alpha\rangle$ and $|\nu_{\rm s}\rangle$ are active and sterile neutrino states, $|\nu_1\rangle$ and $|\nu_2\rangle$ are mass eigenstates of masses m_1 and m_2, where we order $m_1 < m_2$, and θ is the vacuum mixing angle between sterile and active neutrinos. This mixing should be weak, $\theta \ll 1$, otherwise sterile neutrinos would decay too rapidly, see below. The heavy state is mostly sterile neutrinos $|\nu_2\rangle \approx |\nu_{\rm s}\rangle$, and $m_2 \equiv m_{\rm s}$ is the sterile neutrino mass.

The calculation of sterile neutrino abundance is fairly complicated, and we do not reproduce it here. If there is no sizeable lepton asymmetry in the Universe, the sterile neutrino production is most efficient at temperature around

$$T_* \sim \left(\frac{m_{\rm s}}{5G_{\rm F}}\right)^{1/3} \simeq 200\ {\rm MeV} \cdot \left(\frac{m_{\rm s}}{1\ {\rm keV}}\right)^{1/3} .$$

The resulting number density of sterile neutrinos is estimated as

$$\frac{n_{\nu_{\rm s}}}{n_{\nu_\alpha}} \sim T_*^3 M_{\rm Pl}^* G_{\rm F}^2 \cdot \sin^2 2\theta \sim 10^{-2} \cdot \left(\frac{m_{\rm s}}{1\ {\rm keV}}\right) \cdot \left(\frac{\sin 2\theta}{10^{-4}}\right)^2 . \qquad (70)$$

The number density of relic active neutrinos today is about 110 cm^{-3}, so we find from Eq. (70) the estimate for the present contribution of sterile neutrinos into energy density,

$$\Omega_{\nu_s} \simeq 0.2 \cdot \left(\frac{\sin 2\theta}{10^{-4}}\right)^2 \cdot \left(\frac{m_\nu}{1\ \text{keV}}\right)^2 \ . \tag{71}$$

Thus, a sterile neutrino of mass $m_\nu \gtrsim 1$ keV and small mixing angle $\theta_\alpha \lesssim 10^{-4}$ would serve as a dark matter candidate. However, this range of masses and mixing angles is ruled out. The point is that due to its mixing with an active neutrino, a sterile neutrino can decay into an active neutrino and a photon,

$$\nu_s \to \nu_\alpha + \gamma \ .$$

The sterile neutrino decay width is proportional to $\sin^2 2\theta$. If sterile neutrinos are dark matter particles, their decays would produce a narrow line in X-ray flux from the cosmos (orbiting velocity of dark matter particles in our Galaxy is small, $v \sim 10^{-3}$, hence the photons produced in their two-body decays are nearly monochromatic). Such a line has not been observed, and there exist quite strong limits. These limits, translated into limits on $\sin^2 2\theta$ as a function of sterile neutrino mass, are shown in Fig. 12; they rule out the range of masses giving the right mass density of dark matter, Eq. (71). Recall that the mass of a sterile neutrino should exceed 3 keV (in fact, a more precise limit is $m_s > 5.7$ keV [56]).

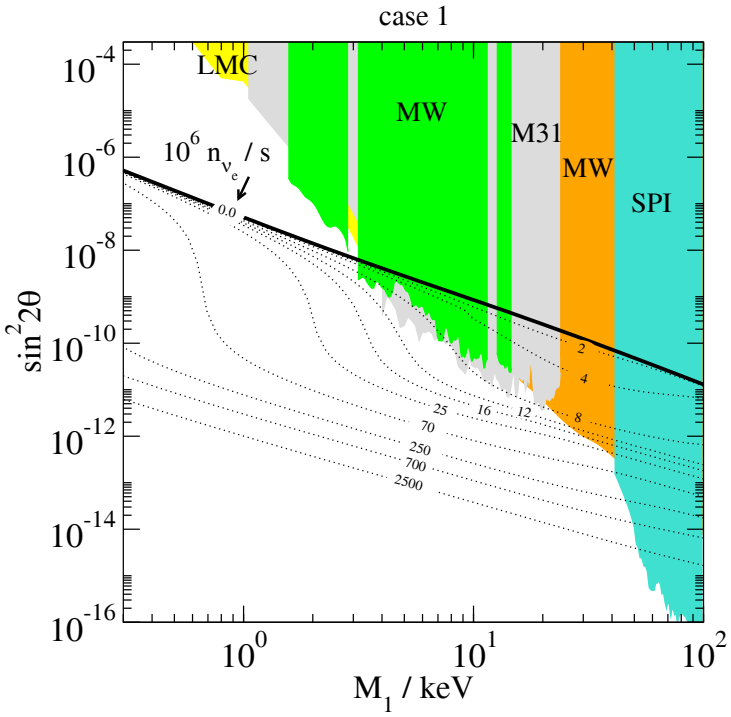


Fig. 12: Limits on sterile neutrino parameters (mass M_1, mixing angle θ) obtained from X-ray telescopes. Solid line corresponds to sterile neutrino dark matter produced in non-resonant oscillations, Eq. (71). Dashed lines show the case of resonant oscillations at non-zero lepton asymmetry; numbers in unit of 10^{-6} show the values of lepton asymmetry (lepton-to-photon ratio η_L) [57].

A (rather baroque) way out [58] is to assume that there is a fairly large lepton asymmetry in the Universe. Then the oscillations of active neutrinos into sterile neutrinos may be enhanced due to the Mikheyev-Smirnov-Wolfenstein (MSW) effect, as at some temperature they occur in the Mikheev–Smirnov resonance regime. In that case the right abundance of sterile neutrinos is obtained at smaller θ, and may be consistent with X-ray bounds. This is also shown in Fig. 12.

4.5.2 Light gravitino

A gravitino—a superpartner of a graviton—is necessarily present in supersymmetric (SUSY) theories. It acquires mass as a result of SUSY breaking (super-Higgs mechanism). The gravitino mass is of order

$$m_{3/2} \simeq \frac{F}{M_{\rm Pl}} \ ,$$

where $\sqrt{F}$ is the supersymmetry breaking scale. Hence, gravitino masses are in the right WDM ballpark for rather low supersymmetry breaking scales, $\sqrt{F} \sim 10^6$–10^7 GeV. This is the case, e.g., in the gauge-mediation scenario. With so low mass, a gravitino is the lightest supersymmetric particle (LSP), so it is stable in many supersymmetric extensions of the Standard Model. From this viewpoint gravitinos can indeed serve as dark matter particles. For what follows, important parameters are the widths of decays of other superpartners into gravitinos and the Standard Model particles. These are of order

$$\Gamma_{\tilde{\rm S}} \simeq \frac{M_{\tilde{\rm S}}^5}{F^2} \simeq \frac{M_{\tilde{\rm S}}^5}{m_{3/2}^2 M_{\rm Pl}^2} \ , \tag{72}$$

where $M_{\tilde{\rm S}}$ is the mass of the superpartner.

One mechanism of the gravitino production in the early Universe is decays of other superpartners. A gravitino interacts with everything else so weakly that once produced, it moves freely, without interacting with cosmic plasma. At production, gravitinos are relativistic and hence they are indeed *warm* dark matter candidates. Let us assume that production in decays is the dominant mechanism and consider under what circumstances the present mass density of gravitinos coincides with that of dark matter.

The rate of gravitino production in decays of superpartners of the type $\tilde{\rm S}$ in the early Universe is

$$\frac{{\rm d}(n_{3/2}/s)}{{\rm d}t} = \frac{n_{\tilde{\rm S}}}{s} \Gamma_{\tilde{\rm S}} \ ,$$

where $n_{3/2}$ and $n_{\tilde{\rm S}}$ are number densities of gravitinos and superpartners, respectively, and s is the entropy density. For superpartners in thermal equilibrium, one has $n_{\tilde{\rm S}}/s = {\rm const} \sim g_*^{-1}$ for $T \gtrsim M_{\tilde{\rm S}}$, and $n_{\tilde{\rm S}}/s \propto \exp(-M_{\tilde{\rm S}}/T)$ at $T \ll M_{\tilde{\rm S}}$. Hence, the production is most efficient at $T \sim M_{\tilde{\rm S}}$, when the number density of superpartners is still large, while the Universe expands most slowly. The density of gravitinos produced in decays of the $\tilde{\rm S}$ is thus given by

$$\frac{n_{3/2}}{s} \simeq \frac{\Gamma_{\tilde{\rm S}}}{g_*} H^{-1}(T \sim M_{\tilde{\rm S}}) \simeq \frac{1}{g_*} \cdot \frac{M_{\tilde{\rm S}}^5}{m_{3/2}^2 M_{\rm Pl}^2} \cdot \frac{M_{\rm Pl}^*}{M_{\tilde{\rm S}}^2} \ .$$

This gives the mass-to-entropy ratio today,

$$\frac{m_{3/2} n_{3/2}}{s} \simeq \sum_{\tilde{\rm S}} \frac{M_{\tilde{\rm S}}^3}{g_*^{3/2} M_{\rm Pl} m_{3/2}} \ , \tag{73}$$

where the sum runs over all superpartner species *which have ever been relativistic in thermal equilibrium*. The correct value (26) is obtained for gravitino masses in the range (29) at

$$M_{\tilde{\rm S}} = 100\text{–}300 \ {\rm GeV} \ . \tag{74}$$

Thus, the scenario with a gravitino as a warm dark matter particle requires light superpartners [59], which are to be discovered at the LHC.

A few comments are in order. First, decays of superpartners is not the only mechanism of gravitino production: gravitinos may also be produced in scattering of superpartners [60]. To avoid overproduction of gravitinos in the latter processes, one has to assume that the maximum temperature in the Universe

(e.g., reached after the post-inflationary reheating stage) is quite low, $T_{\rm max} \sim$ 1–10 TeV. This is not a particularly plausible assumption, but it is consistent with everything else in cosmology and can indeed be realized in some models of inflation. Second, existing constraints on masses of strongly interacting superpartners (gluinos and squarks) suggest that their masses exceed (74). Hence, these particles should not contribute to the sum in (73), otherwise WDM gravitinos would be overproduced. This is possible if masses of squarks and gluinos are larger than $T_{\rm max}$, so that they were never abundant in the early Universe. Third, a gravitino produced in decays of superpartners is *not* a thermal relic, as it was never in thermal equilibrium with the rest of the cosmic plasma. Nevertheless, since gravitinos are produced at $T \sim M_{\tilde{\rm S}}$ and at that time have energy $E \sim M_{\tilde{\rm S}} \sim T$, our estimate (28) does apply.

Question. Let $\tilde{\rm S}$ be the next-to-lightest superpartner which decays into a gravitino of mass $m_{3/2} = 5$ keV and a Standard Model particle. Let $\tilde{\rm S}$ be produced at the LHC at subrelativistic velocity. How far is the decay vertex of $\tilde{\rm S}$ displaced from the proton collision point? Give numerical estimates for $M_{\tilde{\rm S}} =$ 100 GeV and $M_{\tilde{\rm S}} = 1$ TeV.

4.6 Discussion

If dark matter particles are indeed WIMPs, and the relevant energy scale is of order 1 TeV, then the hot Big Bang theory will be probed experimentally up to a temperature of (a few) $\cdot$ (10–100) GeV and down to an age of 10^{-9}–10^{-11} s in the relatively near future (compare to 1 MeV and 1 s accessible today through BBN). With microscopic physics to be known from collider experiments, the WIMP density will be reliably calculated and checked against the data from observational cosmology. Thus, the WIMP scenario offers a window to a very early stage of the evolution of the Universe.

Search for dark matter axions and signals from light sterile neutrinos makes use of completely different methods. Yet there is a good chance for discovery if either of these particles make dark matter.

If dark matter particles are gravitinos, the prospect of probing quantitatively such an early stage of the cosmological evolution is not so bright: it would be very hard, if at all possible, to get an experimental handle on the gravitino mass; furthermore, the present gravitino mass density depends on an unknown reheat temperature $T_{\rm max}$. On the other hand, if this scenario is realized in nature, then the whole picture of the early Universe will be quite different from our best guess on the early cosmology. Indeed, the gravitino scenario requires a low reheat temperature, which in turn calls for a rather exotic mechanism of inflation.

The mechanisms discussed here are by no means the only ones capable of producing dark matter, and the particles we discussed are by no means the only candidates for dark matter particles. Other dark matter candidates include axinos, Q-balls, very heavy relics produced towards the end of inflation (wimpzillas) etc. Hence, even though there are grounds to hope that the dark matter problem will be solved soon, there is no guarantee at all.

5 Baryon asymmetry of the Universe

As we discussed in Section 2.4, the baryon asymmetry of the Universe is characterized by the baryon-to-entropy ratio, which at high temperatures is defined as follows:

$$\Delta_{\rm B} = \frac{n_{\rm B} - n_{\bar{\rm B}}}{s} \,,$$

where $n_{\bar{\rm B}}$ is the number density of antibaryons and s is the entropy density. If the baryon number is conserved and the Universe expands adiabatically (which is the case at least after the electroweak epoch, $T \lesssim 100$ GeV), $\Delta_{\rm B}$ is time-independent and equal to its present value $\Delta_{\rm B} \approx 0.8 \times 10^{-10}$, see Eq. (25). At early times, at temperatures well above 100 MeV, cosmic plasma contained many quark–antiquark

pairs, whose number density was of the order of the entropy density,

$$n_{\rm q} + n_{\bar{\rm q}} \sim s \ ,$$

while the baryon number density was related to densities of quarks and antiquarks as follows (baryon number of a quark equals $1/3$):

$$n_{\rm B} = \frac{1}{3}(n_{\rm q} - n_{\bar{\rm q}}) \ .$$

Hence, in terms of quantities characterizing the very early epoch, the baryon asymmetry may be expressed as

$$\Delta_{\rm B} \sim \frac{n_{\rm q} - n_{\bar{\rm q}}}{n_{\rm q} + n_{\bar{\rm q}}} \ .$$

We see that there was one extra quark per about 10 billion quark–antiquark pairs. It is this tiny excess that is responsible for the entire baryonic matter in the present Universe: as the Universe expanded and cooled down, antiquarks annihilated with quarks, and only the excessive quarks remained and formed baryons.

There is no logical contradiction to suppose that the tiny excess of quarks over antiquarks was built in as an initial condition. This is not at all satisfactory for a physicist, however. Furthermore, the inflationary scenario does not provide such an initial condition for the hot Big Bang epoch; rather, inflationary theory predicts that the Universe was baryon-symmetric just after inflation. Hence, one would like to explain the baryon asymmetry dynamically [61, 62], i.e., find the mechanism of its generation in the early Universe.

5.1 Sakharov conditions

The baryon asymmetry may be generated from an initially baryon-symmetric state only if three necessary conditions, dubbed Sakharov conditions, are satisfied. These are:

1. baryon number non-conservation;
2. C- and CP-violation;
3. deviation from thermal equilibrium.

All three conditions are easily understood. (1) If baryon number were conserved, and initial net baryon number in the Universe was zero, the Universe today would still be symmetric. (2) If C or CP were conserved, then the rate of reactions with particles would be the same as the rate of reactions with antiparticles, and no asymmetry would be generated. (3) Thermal equilibrium means that the system is stationary (no time dependence at all). Hence, if the initial baryon number is zero, it is zero forever, unless there are deviations from thermal equilibrium. Furthermore, if there are processes that violate baryon number, and the system approaches thermal equilibrium, then the baryon number tends to be washed out rather than generated.

At the epoch of the baryon-asymmetry generation, all three Sakharov conditions have to be met simultaneously. There is a qualification, however. These conditions would be literally correct if there were no other relevant quantum numbers that characterize the cosmic medium. In reality, however, lepton numbers also play a role. As we will see shortly, baryon and lepton numbers are rapidly violated by anomalous electroweak processes at temperatures above, roughly, 100 GeV. What is conserved in the Standard Model is the combination $B - L$, where L is the total lepton number. So, there are two options. One is to generate the baryon asymmetry at or below the electroweak epoch, $T \lesssim 100$ GeV, and make sure that the electroweak processes do not wash out the baryon asymmetry after its generation. This leads to the idea of electroweak baryogenesis (another possibility is Affleck–Dine baryogenesis [63]). Another is to generate $B - L$ asymmetry before the electroweak epoch, i.e., at $T \gg 100$ GeV: if the Universe

is $B - L$ asymmetric above 100 GeV, the electroweak physics reprocesses $B - L$ partially into baryon number and partially into lepton number, so that in thermal equilibrium with conserved $B - L$ one has

$$B = C \cdot (B - L) \,, \qquad L = (C - 1) \cdot (B - L) \,,$$

where C is a constant of order 1 ($C = 28/79$ in the Standard Model at $T \gtrsim 100$ GeV). In the second scenario, the first Sakharov condition applies to $B - L$ rather than baryon number itself.

Let us point out two most common mechanisms of baryon number non-conservation. One emerges in grand unified theories and is due to the exchange of supermassive particles. It is similar, say, to the mechanism of charm non-conservation in weak interactions, which occurs via the exchange of heavy W bosons. The scale of these new, baryon number violating interactions is the grand unification scale, presumably of order $M_{\rm GUT} \simeq 10^{16}$ GeV. It is rather unlikely, however, that the baryon asymmetry was generated due to this mechanism: the relevant temperature would be of order $M_{\rm GUT}$, while such a high reheat temperature after inflation is difficult to obtain.

Another mechanism is non-perturbative [39] and is related to the triangle anomaly in the baryonic current (a keyword here is 'sphaleron' [64,65]). It exists already in the Standard Model and, possibly with mild modifications, operates in all its extensions. The two main features of this mechanism, as applied to the early Universe, is that it is effective over a wide range of temperatures, 100 GeV $< T < 10^{11}$ GeV, and, as we pointed out above, that it conserves $B - L$.

5.2 Electroweak baryon number non-conservation

Let us pause here to discuss the physics behind electroweak baryon and lepton number non-conservation in a little more detail, though still at a qualitative level. A detailed analysis can be found in the book [66] and in references therein.

Let us consider the baryonic current,

$$B^\mu = \frac{1}{3} \cdot \sum_i \bar{q}_i \gamma^\mu q_i \,,$$

where the sum runs over quark flavours. Naively, it is conserved, but at the quantum level its divergence is non-zero because of the triangle anomaly (a similar effect goes under the name of the axial anomaly in the context of quantum electrodynamics (QED) and QCD),

$$\partial_\mu B^\mu = \frac{1}{3} \cdot 3_{\rm colours} \cdot 3_{\rm generations} \cdot \frac{g^2}{32\pi^2} \epsilon^{\mu\nu\lambda\rho} F^{\rm a}_{\mu\nu} F^{\rm a}_{\lambda\rho} \,,$$

where $F^{\rm a}_{\mu\nu}$ and g are the field strength of the SU(2)$_{\rm W}$ gauge field and the SU(2)$_{\rm W}$ gauge coupling, respectively. Likewise, each leptonic current ($\alpha = {\rm e}, \mu, \tau$) is anomalous in the Standard Model (we disregard here neutrino masses and mixings, which violate lepton numbers too),

$$\partial_\mu \mathcal{L}^\mu_\alpha = \frac{g^2}{32\pi^2} \cdot \epsilon^{\mu\nu\lambda\rho} F^{\rm a}_{\mu\nu} F^{\rm a}_{\lambda\rho} \,. \tag{75}$$

A non-trivial fact is that there exist large field fluctuations, $F^{\rm a}_{\mu\nu}(\mathbf{x}, t) \propto g^{-1}$, such that

$$Q \equiv \int {\rm d}^3x {\rm d}t \, \frac{g^2}{32\pi^2} \cdot \epsilon^{\mu\nu\lambda\rho} F^{\rm a}_{\mu\nu} F^{\rm a}_{\lambda\rho} \neq 0 \,. \tag{76}$$

Furthermore, for any physically relevant fluctuation the value of Q is an integer ('physically relevant' means that the gauge field strength vanishes at infinity in space–time). In four space–time dimensions such fluctuations exist only in *non-Abelian* gauge theories.

Suppose now that a fluctuation with non-vanishing Q has occurred. Then the baryon numbers at the end and beginning of the process are different,

$$B_{\rm fin} - B_{\rm in} = \int {\rm d}^3x{\rm d}t\, \partial_\mu B^\mu = 3Q \ . \tag{77}$$

Likewise,

$$\mathcal{L}\alpha,\ {\rm fin} - \mathcal{L}_{\alpha,\ {\rm in}} = Q \ . \tag{78}$$

This explains the selection rule mentioned above: B is violated, $B - L$ is not.

At zero temperature, the field fluctuations that induce baryon and lepton number violation are vacuum fluctuations, called instantons [67]. Since these are *large* field fluctuations, their probability is exponentially suppressed. The suppression factor in the Standard Model is

$$\mathrm{e}^{-\frac{16\pi^2}{g^2}} \sim 10^{-165} \ .$$

Therefore, the rate of baryon number violating processes at zero temperature is suppressed by this factor, making these processes totally negligible. On the other hand, at high temperatures there are large *thermal* fluctuations ('sphalerons') whose rate is not necessarily small. And, indeed, B-violation in the early Universe is rapid as compared to the cosmological expansion at sufficiently high temperatures, provided that (see Ref. [11] for details)

$$\langle\phi\rangle_T < T \ , \tag{79}$$

where $\langle\phi\rangle_T$ is the Englert–Brout–Higgs expectation value at temperature T.

One may wonder how baryon number is not conserved in the absence of explicit baryon number violating terms in the Lagrangian of the Standard Model. To understand what is going on, let us consider a massless *left-handed* fermion field in the background of the SU(2) gauge field $\mathbf{A}(\mathbf{x}, t)$, which depends on space–time coordinates in a non-trivial way. As a technicality, we set the temporal component of the gauge field equal to zero, $A_0 = 0$, by the choice of gauge. One way to understand the behaviour of the fermion field in the gauge field background is to study the system of eigenvalues of the Dirac Hamiltonian $\{\omega(t)\}$. The Hamiltonian is defined in the standard way

$$H_{\rm Dirac}(t) = {\rm i}\alpha^i \left(\partial_i - {\rm i}gA_i(\mathbf{x}, t)\right) \frac{1-\gamma_5}{2} \ ,$$

where $\alpha^i = \gamma^0\gamma^i$, so that the Dirac equation has the Schrödinger form,

$${\rm i}\frac{\partial\psi}{\partial t} = H_{\rm Dirac}\psi \ .$$

So, let us discuss the eigenvalues $\omega_n(t)$ of the operator $H_{\rm Dirac}(t)$, treating t as a parameter. These eigenvalues are found from

$$H_{\rm Dirac}(t)\psi_n = \omega_n(t)\psi_n \ .$$

At $\mathbf{A} = 0$, the system of levels is shown schematically in Fig. 13. Importantly, there are both positive- and negative-energy levels. According to Dirac, the lowest-energy state (Dirac vacuum) has all negative-energy levels occupied, and all positive-energy levels empty. Occupied positive-energy levels (three of them in Fig. 13) correspond to real fermions, while empty negative-energy levels describe antifermions (one in Fig. 13). Fermion–antifermion annihilation in this picture is a jump of a fermion from a positive-energy level to an unoccupied negative-energy level. As a side remark, this original Dirac picture is, in fact, equivalent to the more conventional (by now) procedure of the quantization of the fermion field, which does not make use of the notion of negative-energy levels. The discussion that follows can be translated into the conventional language; however, the original Dirac picture turns out to be a lot more

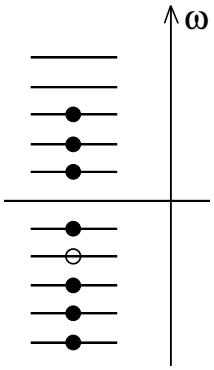


Fig. 13: Fermion energy levels at zero background gauge field

transparent in our context. This is a nice example of the complementarity of various approaches in quantum field theory.

Let us proceed with the discussion of the fermion energy levels in gauge field backgrounds. In weak background fields, the energy levels depend on time ('move'), but nothing dramatic happens. For adiabatically varying background fields, the fermions merely sit on their levels, while fast-changing fields generically give rise to jumps from, say, negative- to positive-energy levels, that is, creation of fermion–antifermion pairs. Needless to say, fermion number $(N_{\rm f} - N_{\bar{\rm f}})$ is conserved.

The situation is entirely different for the background fields with non-zero Q. The levels of left-handed fermions move as shown in the left-hand panel of Fig. 14. Some levels necessarily cross zero, and

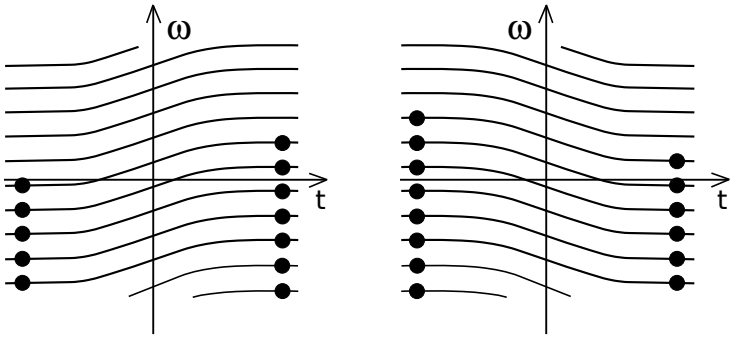


Fig. 14: Motion of fermion levels in background gauge fields with non-vanishing Q (shown is the case $Q = 2$). Left-hand panel: left-handed fermions. Right-hand panel: right-handed fermions.

the net number of levels crossing zero from below equals Q. This means that the number of left-handed fermions is not conserved: for an adiabatically varying gauge field $\mathbf{A}(\mathbf{x}, t)$, the motion of levels shown in the left-hand panel of Fig. 14 corresponds to the case in which the initial state of the fermionic system is vacuum (no real fermions or antifermions) whereas the final state contains Q real fermions (two in the particular case shown). If the evolution of the gauge field is not adiabatic, the result for the fermion number non-conservation is the same: there may be jumps from negative-energy levels to positive-energy levels or vice versa. These correspond to creation or annihilation of fermion–antifermion pairs, but the net change of the fermion number (number of fermions minus number of antifermions) remains equal to Q. Importantly, the initial and final field configurations of the gauge field may be trivial, $\mathbf{A} = 0$ (up to gauge transformation), so that fermion number non-conservation may occur due to a fluctuation that begins and ends in the gauge field vacuum. These are precisely instanton-like vacuum fluctuations. At finite temperatures, processes of this type occur due to thermal fluctuations, i.e., sphalerons.

If the same gauge field interacts also with right-handed fermions, the motion of the levels of the latter is opposite to that of left-handed fermions. This is shown in the right-hand panel of Fig. 14. The change in the number of right-handed fermions is equal to $-Q$. So, if the gauge interaction is vector-like, the total fermion number $(N_{\rm left} + N_{\rm right})$ is conserved, while chirality $(N_{\rm left} - N_{\rm right})$ is violated even for massless fermions. This explains why there is no baryon number violation in QCD. The above discussion implies, instead, that there is non-perturbative violation of chirality in QCD in the limit of massless quarks. The latter phenomenon has non-trivial consequences, which are indeed confirmed by phenomenology. In this sense anomalous non-conservation of fermion quantum numbers is an experimentally established fact.

In electroweak theory, right-handed fermions do not interact with the $\mathrm{SU}(2)_{\rm W}$ gauge field, while left-handed fermions do. Therefore, fermion number is not conserved (the anomalous relations (75) and (76) suggest that this result is valid also in the presence of the Standard Model Yukawa couplings of quarks and leptons; this is indeed the case). Since fermions of each $\mathrm{SU}(2)_{\rm W}$ doublet interact with the $\mathrm{SU}(2)_{\rm W}$ gauge bosons in one and the same way, they are equally created in a process involving a gauge field fluctuation with non-zero Q. This again leads to the relations (77) and (78), i.e., to the selection rules $\Delta B = \Delta L$ and $\Delta L_{\rm e} = \Delta L_\mu = \Delta L_\tau$.

5.3 Electroweak baryogenesis

It is tempting to make use of the electroweak mechanism of baryon number non-conservation for explaining the baryon asymmetry of the Universe. This scenario is known as electroweak baryogenesis. It meets two problems, however. One is that CP-violation in the Standard Model is too weak: the CKM mechanism alone is insufficient to generate a realistic value of the baryon asymmetry. Hence, one needs extra sources of CP-violation. Another problem has to do with departure from thermal equilibrium that is necessary for the generation of the baryon asymmetry. At temperatures well above 100 GeV, electroweak symmetry is restored, the expectation value of the Englert–Brout–Higgs field ϕ is zero, the relation (79) is valid and the baryon number non-conservation is rapid as compared to the cosmological expansion. At temperatures of order 100 GeV, the relation (79) may be violated, but the Universe expands very slowly: the cosmological time-scale at these temperatures is

$$H^{-1} = \frac{M^*_{\rm Pl}}{T^2} \sim 10^{-10}\ {\rm s}\ , \tag{80}$$

which is very large by the electroweak physics standards. The only way in which a strong departure from thermal equilibrium at these temperatures may occur appears to be the first-order phase transition.

The property that at temperatures well above 100 GeV the expectation value of the Englert–Brout–Higgs field is zero, while it is non-zero in vacuo, suggests that there may be a phase transition from the phase with $\langle\phi\rangle = 0$ to the phase with $\langle\phi\rangle \neq 0$. The situation is pretty subtle here, as ϕ is not gauge invariant and hence cannot serve as an order parameter, so the notion of phases with $\langle\phi\rangle = 0$ and $\langle\phi\rangle \neq 0$ is vague. In fact, neither electroweak theory nor most of its extensions have a gauge-invariant order parameter, so there is no real distinction between these 'phases'. This situation is similar to that in a liquid–vapour system, which does not have an order parameter and may or may not experience a vapour–liquid phase transition as temperature decreases, depending on other, external parameters characterizing this system, e.g., pressure. In the Standard Model the role of such an 'external' parameter is played by the Englert–Brout–Higgs self-coupling λ or, in other words, the Higgs boson mass.

Continuing to use somewhat sloppy terminology, we recall that in thermal equilibrium any system is at the global minimum of its *free energy*. To figure out the expectation value of ϕ at a given temperature, one introduces the temperature-dependent effective potential $V_{\rm eff}(\phi; T)$, which is equal to the free energy density in the system where the average field is pinpointed to a prescribed value ϕ, but otherwise there is thermal equilibrium. Then the global minimum of $V_{\rm eff}$ at a given temperature is at the equilibrium value of ϕ, while local minima correspond to metastable states.

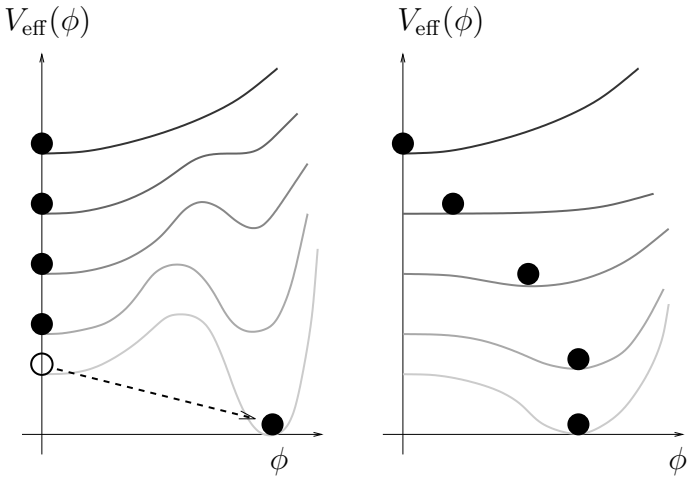


Fig. 15: Effective potential as function of ϕ at different temperatures. Left: first-order phase transition. Right: second-order phase transition. Upper curves correspond to higher temperatures. Black blobs show the expectation value of ϕ in thermal equilibrium. The arrow in the left-hand panel illustrates the transition from the metastable, supercooled state to the ground state.

The interesting case for us is the first-order phase transition. In this case, the system evolves as follows. At high temperatures, there exists one minimum of $V_{\rm eff}$ at $\phi = 0$, and the expectation value of the Englert–Brout–Higgs field is zero. As the temperature decreases, another minimum appears at finite ϕ, and then becomes lower than the minimum at $\phi = 0$; see left-hand panel of Fig. 15. However, the minima with $\phi = 0$ and $\phi \neq 0$ are separated by a barrier of $V_{\rm eff}$, the probability of the transition from the phase $\phi = 0$ to the phase $\phi \neq 0$ is very small for some time and the system gets overcooled. The transition occurs when the temperature becomes sufficiently low and the transition probability sufficiently high. This is to be contrasted to the case, e.g., of the second-order phase transition, right-hand panel of Fig. 15. In the latter case, the field slowly evolves, as the temperature decreases, from zero to non-zero vacuum value, and the system remains very close to thermal equilibrium at all times.

During the first-order phase transition, the field cannot jump from $\phi = 0$ to $\phi \neq 0$ homogeneously throughout the whole space: intermediate homogeneous configurations have free energies proportional to the volume of the system (recall that $V_{\rm eff}$ is free energy *density*), i.e., infinite. Instead, the transition occurs just like the first-order vapour–liquid transition, through boiling. Thermal fluctuations spontaneously create bubbles of the new phase inside the old phase. These bubbles then grow, their walls eventually collide and the new phase finally occupies the entire space. The Universe boils. In the cosmological context, this process happens when the bubble nucleation rate per Hubble time per Hubble volume is roughly of order 1, i.e., when a few bubbles are created in Hubble volume in Hubble time. The velocity of the bubble wall in the relativistic cosmic plasma is roughly of the order of the speed of light (in fact, it is somewhat smaller, from 0.1 to 0.01), simply because there are no relevant dimensionless parameters characterizing the system. Hence, the bubbles grow large before their walls collide: their size at collision is roughly of the order of the Hubble size (in fact, one or two orders of magnitude smaller). While the bubble is microscopic at nucleation—its size is determined by the electroweak scale and is roughly of order $(100\ \text{GeV})^{-1} \sim 10^{-16}$ cm—its size at collision of walls is macroscopic, $R \sim 10^{-2}$–10^{-3} cm, as follows from (80). Clearly, boiling is a highly non-equilibrium process, and one may hope that the baryon asymmetry may be generated at that time. And, indeed, there exist mechanisms of the generation of the baryon asymmetry, which have to do with interactions of quarks and leptons with moving bubble walls. The value of the resulting baryon asymmetry may well be of order 10^{-10}, as required by observations, provided that there is enough CP-violation in the theory.

A necessary condition for the electroweak generation of the baryon asymmetry is that the inequality (79) must be violated *just after* the phase transition. Indeed, in the opposite case the electroweak

baryon number violating processes are fast after the transition, and the baryon asymmetry, generated during the transition, is washed out afterwards. Hence, the phase transition must be of strong enough first order. This is *not* the case in the Standard Model. To see why this is so, and to get an idea of in which extensions of the Standard Model the phase transition may be of strong enough first order, let us consider the effective potential in some detail. At zero temperature, the Englert–Brout–Higgs potential has the standard form,

$$V(\phi) = -\frac{m^2}{2}|\phi|^2 + \frac{\lambda}{4}|\phi|^4 \ .$$

Here

$$|\phi| \equiv \left(\phi^\dagger \phi\right)^{1/2} \tag{81}$$

is the length of the Englert–Brout–Higgs doublet ϕ, $m^2 = \lambda v^2$ and $v = 246$ GeV is the Englert–Brout–Higgs expectation value in vacuo. The Higgs boson mass is related to the latter as follows:

$$m_{\rm H} = \sqrt{2\lambda} v \ . \tag{82}$$

Now, to the leading order of perturbation theory, the finite-temperature effects modify the effective potential into

$$V_{\rm eff}(\phi, T) = \frac{\alpha(T)}{2}|\phi|^2 - \frac{\beta}{3}T|\phi|^3 + \frac{\lambda}{4}|\phi|^4 \ . \tag{83}$$

Here $\alpha(T) = -m^2 + \hat{g}^2 T^2$, where $\hat{g}^2$ is a positive linear combination of squares of coupling constants of all fields to the Englert–Brout–Higgs field (in the Standard Model, a linear combination of g^2, g'^2 and y_i^2, where g and g' are $\mathrm{SU}(2)_{\rm W}$ and $\mathrm{U}(1)_{\rm Y}$ gauge couplings and y_i are Yukawa couplings). The phase transition occurs roughly when $\alpha(T) = 0$. An important parameter β is a positive linear combination of cubes of coupling constants of all *bosonic* fields to the Englert–Brout–Higgs field. In the Standard Model, β is a linear combination of g^3 and g'^3, i.e., a linear combination of $M_{\rm W}^3/v^3$ and $M_{\rm Z}^3/v^3$,

$$\beta = \frac{1}{2\pi}\frac{2M_{\rm W}^3 + M_{\rm Z}^3}{v^3} \ . \tag{84}$$

The cubic term in (83) is rather peculiar: in view of (81), it is not analytic in the original Englert–Brout–Higgs field ϕ. Yet this term is crucial for the first-order phase transition: for $\beta = 0$ the phase transition would be of the second order.

Question. Show that the phase transition is second order for $\beta = 0$.

The origin of the non-analytic cubic term can be traced back to the enhancement of the Bose–Einstein thermal distribution at low momenta, $p, m \ll T$,

$$f_{\rm Bose}(p) = \frac{1}{{\rm e}^{\frac{\sqrt{p^2+m_a^2}}{T}} - 1} \simeq \frac{T}{\sqrt{p^2 + m_a^2}} \ ,$$

where $m \simeq g_a|\phi|$ is the mass of the boson a that is generated due to the non-vanishing Englert–Brout–Higgs field, and g_a is the coupling constant of the field a to the Englert–Brout–Higgs field. Clearly, at $p \ll g|\phi|$ the distribution function is non-analytic in ϕ,

$$f_{\rm Bose}(p) \simeq \frac{T}{g_a|\phi|} \ .$$

It is this non-analyticity that gives rise to the non-analytic cubic term in the effective potential. Importantly, the Fermi–Dirac distribution,

$$f_{\rm Fermi}(p) = \frac{1}{{\rm e}^{\frac{\sqrt{p^2+m_a^2}}{T}} + 1} \ ,$$

is analytic in m_a^2, and hence $\phi^\dagger\phi$, so fermions do not contribute to the cubic term.

With the cubic term in the effective potential, the phase transition is indeed of the first order: at high temperatures the coefficient α is positive and large, and there is one minimum of the effective potential at $\phi = 0$, while for α small but still positive there are two minima. The phase transition occurs at $\alpha \approx 0$; at that moment

$$V_{\rm eff}(\phi, T) \approx -\frac{\beta T}{3}|\phi|^3 + \frac{\lambda}{4}|\phi|^4 \; .$$

We find from this expression that immediately after the phase transition the minimum of $V_{\rm eff}$ is at

$$\phi \simeq \frac{\beta T}{\lambda} \; .$$

Hence, the necessary condition for successful electroweak baryogenesis, $\phi > T$, translates into

$$\beta > \lambda \; . \tag{85}$$

According to (82), λ is proportional to $m_{\rm H}^2$, whereas in the Standard Model β is proportional to $(2M_{\rm W}^3 + M_{\rm Z}^3)$. Therefore, the relation (85) holds for small Higgs boson masses only; in the Standard Model one makes use of (82) and (84) and finds that this happens for $m_{\rm H} < 50$ GeV, while in reality $m_{\rm H} = 125$ GeV. In fact, in the Standard Model with $m_{\rm H} = 125$ GeV, there is no phase transition at all; the electroweak transition is a smooth crossover instead. The latter fact is not visible from the expression (83), but that expression is the lowest-order perturbative result, while the perturbation theory is not applicable for describing the transition in the Standard Model with large $m_{\rm H}$.

This discussion indicates a possible way to make the electroweak phase transition strong. What one needs is the existence of new bosonic fields that have large enough couplings to the Englert–Brout–Higgs field(s), and hence provide large contributions to β. To have an effect on the dynamics of the transition, the new bosons must be present in the cosmic plasma at the transition temperature, $T \sim 100$ GeV, so their masses should not be too high, $M \lesssim 300$ GeV. In supersymmetric extensions of the Standard Model, the natural candidate for a long time has been stop (superpartner of top-quark) whose Yukawa coupling to the Englert–Brout–Higgs field is the same as that of top, that is, large. The light stop scenario for electroweak baryogenesis would indeed work, as has been shown by the detailed analysis in Refs. [68–70].

There are other possibilities to make the electroweak transition strongly first order. Generically, they require an extension of the scalar sector of the Standard Model and predict new fairly light scalars which interact with the Standard Model Englert–Brout–Higgs field and may or may not participate in gauge interactions.

Yet another issue is CP-violation, which has to be strong enough for successful electroweak baryogenesis. As the asymmetry is generated in the interactions of quarks and leptons with the bubble walls, CP-violation must occur at the walls. Recall now that the walls are made of the scalar field(s). This points towards the necessity of CP-violation in the scalar sector, which may only be the case in a theory containing scalar fields other than the Standard Model Englert–Brout–Higgs field.

To summarize, electroweak baryogenesis requires a considerable extension of the Standard Model, with masses of new particles in the range 100–300 GeV. Hence, this mechanism will most likely be ruled out or confirmed by the LHC. We emphasize, however, that electroweak baryogenesis is not the only option at all: an elegant and well-motivated competitor is leptogenesis [14, 15, 71]; there are many other mechanisms proposed in the literature.

6 Before the hot epoch

6.1 Cosmological perturbations: preliminaries

With BBN theory and observations, and due to evidence, albeit indirect, for relic neutrinos, we are confident of the theory of the early Universe at temperatures up to $T \simeq 1$ MeV, which correspond to an

age of $t \simeq 1$ s. With the LHC, we hope to be able to learn the Universe up to temperature $T \sim 100$ GeV and age $t \sim 10^{-10}$ s. Are we going to have a handle on an even earlier epoch?

The key issue in this regard is cosmological perturbations. These are inhomogeneities in the energy density and associated gravitational potentials, in the first place. This type of inhomogeneities is called scalar perturbations, as they are described by 3-scalars. There may exist perturbations of another type, called tensors; these are primordial gravity waves. We will mostly concentrate on scalar perturbations, since they are observed; tensor perturbations are important too, and we comment on them later on. While perturbations of the present size of order 10 Mpc and smaller have large amplitudes today and are non-linear, amplitudes of all known perturbations were small in the past, and the perturbations can be described within the linearized theory. Indeed, CMB temperature anisotropy tells us that the perturbations at the recombination epoch were roughly at the level

$$\delta \equiv \frac{\delta\rho}{\rho} = 10^{-4}\text{–}10^{-5} \; . \tag{86}$$

Thus, the linearized theory works very well before recombination and somewhat later. We will be rather sloppy when talking about scalar perturbations. In general relativity, there is arbitrariness in the choice of reference frame, which can be viewed as a sort of gauge freedom. In a homogeneous and isotropic Universe, there is a preferred reference frame, in which quantities like energy density or distribution function of CMB photons are manifestly homogeneous and isotropic. It is in this frame that the metric has FLRW form (1). Once there are perturbations, no preferred reference frame exists any longer. As an example, one can choose a reference frame such that the three-dimensional hypersurfaces of constant time are hypersurfaces of constant total energy density ρ. In this frame one has $\delta\rho = 0$, so Eq. (86) does not make sense. Yet the Universe is inhomogeneous in this reference frame, since there are inhomogeneous metric perturbations $\delta g_{\mu\nu}(\mathbf{x}, t)$. We will skip these technicalities and denote the scalar perturbation by δ without specifying its gauge-invariant meaning.

Equations for perturbations are obtained by writing for every variable (including metric) an expression like $\rho(\mathbf{x}, t) = \bar{\rho}(t) + \delta\rho(\mathbf{x}, t)$ etc, where $\bar{\rho}(t)$ is the homogeneous and isotropic background, which we discussed in Section 2.3. One inserts the perturbed variables into the Einstein equations and covariant conservation equations $\nabla_\mu T^{\mu\nu} = 0$ and linearizes this set of equations. In many cases one also has to use the linearized Boltzmann equations that govern the distribution functions of particles out of thermal equilibrium; these are necessary for evaluating the linearized perturbations of the energy–momentum tensor. In any case, since the background FLRW metric (1) does not explicitly depend on $\mathbf{x}$, the linearized equations for perturbations do not contain $\mathbf{x}$ explicitly. Therefore, one makes use of the spatial Fourier decomposition

$$\delta(\mathbf{x}, t) = \int \mathrm{e}^{\mathrm{i}\mathbf{k}\mathbf{x}} \delta(\mathbf{k}, t) \, \mathrm{d}^3 k \; .$$

The advantage is that modes with different momenta $\mathbf{k}$ evolve independently in the linearized theory, i.e., each mode can be treated separately. Recall that $\mathrm{d}\mathbf{x}$ is *not* the physical distance between neighbouring points; the physical distance is $a(t)\mathrm{d}\mathbf{x}$. Thus, $\mathbf{k}$ is *not* the physical momentum (wavenumber); the physical momentum is $\mathbf{k}/a(t)$. While for a given mode the comoving (or coordinate) momentum $\mathbf{k}$ remains constant in time, the physical momentum gets red shifted as the Universe expands, see also Section 2.1. In what follows we set the present value of the scale factor equal to 1 (in a spatially flat Universe this can always be done by rescaling the coordinates $\mathbf{x}$):

$$a_0 \equiv a(t_0) = 1 \; ;$$

then $\mathbf{k}$ is the *present* physical momentum and $2\pi/k$ is the present physical wavelength, which is also called the comoving wavelength.

Properties of scalar perturbations are measured in various ways. Perturbations of fairly large spatial scales (fairly low $\mathbf{k}$) give rise to CMB temperature anisotropy and polarization, so we have very

detailed knowledge of them. Somewhat shorter wavelengths are studied by analysing distributions of galaxies and quasars at present and in relatively near past. There are several other methods, some of which can probe even shorter wavelengths. There is good overall consistency of the results obtained by different methods, so we have a pretty good understanding of many aspects of the scalar perturbations.

The cosmic medium in our Universe has several components that interact only gravitationally: baryons, photons, neutrinos and dark matter. Hence, there may be and, in fact, there are perturbations in each of these components. As we pointed out in Section 4, electromagnetic interactions between baryons, electrons and photons were strong before recombination, so to a reasonable approximation these species made a single fluid, and it is appropriate to talk about perturbations in this fluid. After recombination, baryons and photons evolved independently.

By studying the scalar perturbations, we have learned a number of very important things. To appreciate what they are, it is instructive to consider first the baryon–electron–photon fluid before recombination.

6.2 Perturbations in the expanding Universe: subhorizon and superhorizon regimes.

Perturbations in the baryon–photon fluid before recombination are nothing but sound waves. It is instructive to compare the wavelength of a perturbation with the horizon size. To this end, recall (see Section 2.4) that the horizon size $l_{\rm H}(t)$ is the size of the largest region which is causally connected by the time t, and that

$$l_{\rm H}(t) \sim H^{-1}(t) \sim t$$

at radiation domination and later, see Eq. (21). The latter relation, however, holds *under the assumption that the hot epoch was the first one in cosmology*, i.e., that the radiation domination started right after the Big Bang. This assumption is in the heart of what can be called hot Big Bang theory. We will find that this assumption in fact is *not valid* for our Universe; we are going to see this ad absurdum, so let us stick to the hot Big Bang theory for the time being.

Unlike the horizon size, the physical wavelength of a perturbation grows more slowly. As an example, at radiation domination

$$\lambda(t) = \frac{2\pi a(t)}{k} \propto \sqrt{t}\,,$$

while at matter domination $\lambda(t) \propto t^{2/3}$. For obvious reasons, the modes with $\lambda(t) \ll H^{-1}(t)$ and $\lambda(t) \gg H^{-1}(t)$ are called subhorizon and superhorizon at time t, respectively. We are able to study the modes which are subhorizon *today*; longer modes are homogeneous throughout the visible Universe and are not observed as perturbations. However, *the wavelengths which are subhorizon today were superhorizon at some earlier epoch*. In other words, the physical momentum $k/a(t)$ was smaller than $H(t)$ at early times; at time $t_\times$ such that

$$q(t_\times) \equiv \frac{k}{a(t_\times)} = H(t_\times)\,,$$

the mode entered the horizon, and after that evolved in the subhorizon regime $k/a(t) \gg H(t)$, see Fig. 16. It is straightforward to see that for all cosmologically interesting wavelengths, horizon crossing occurs much later than 1 s after the Big Bang, i.e., at the time we are confident about. So, there is no guesswork at this point.

Question. Estimate the temperature at which a perturbation of comoving size 10 kpc entered the horizon.

Another way to look at the superhorizon–subhorizon behaviour of perturbations is to introduce a

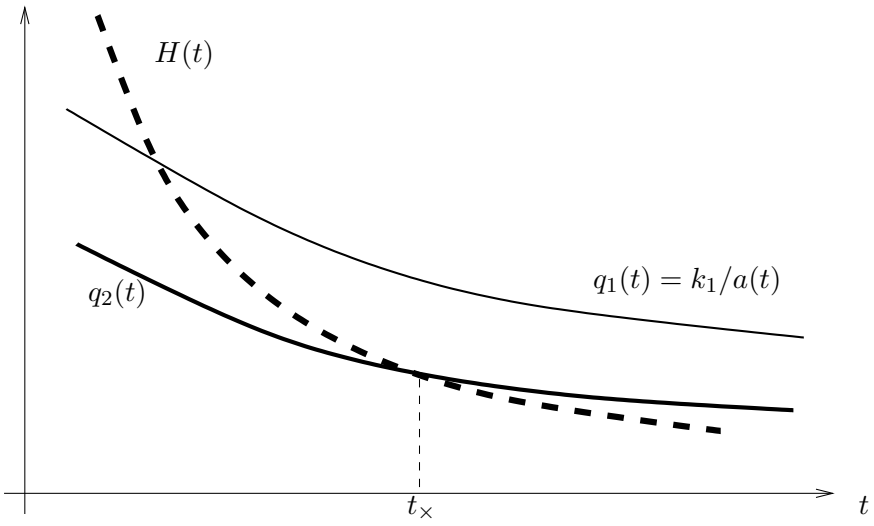


Fig. 16: Physical momenta $q(t) = k/a(t)$ (solid lines, $k_2 < k_1$) and Hubble parameter (dashed line) at radiation- and matter-dominated epochs. Here $t_\times$ is the horizon entry time.

new time coordinate (cf. Eq. (20)),

$$\eta = \int_0^t \frac{\mathrm{d}t'}{a(t')} \ . \tag{87}$$

Note that this integral converges at the lower limit in the hot Big Bang theory. In terms of this time coordinate, the FLRW metric (1) reads

$$\mathrm{d}s^2 = a^2(\eta)(\mathrm{d}\eta^2 - \mathrm{d}\mathbf{x}^2) \ .$$

In coordinates $(\eta, \mathbf{x})$, the light cones $\mathrm{d}s = 0$ are the same as in Minkowski space, and η is the coordinate size of the horizon, see Fig. 17.

Every mode of perturbation has a time-independent coordinate wavelength $2\pi/k$, and at small η it is in superhorizon regime, $2\pi/k \gg \eta$, and after horizon crossing at time $\eta_\times = \eta(t_\times)$ it becomes subhorizon.

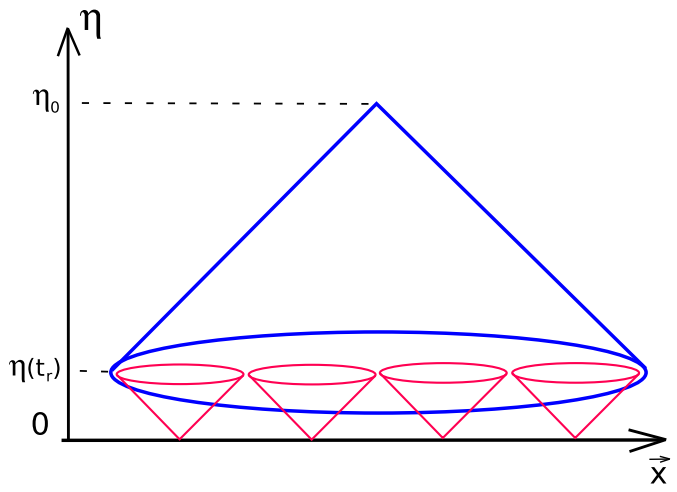


Fig. 17: Causal structure of space–time in the hot Big Bang theory. Here $\mathrm{t_r}$ is the conformal time at recombination.

6.3 Hot epoch was not the first

One immediately observes that this picture falsifies the hot Big Bang theory. Indeed, we see the horizon at recombination $l_{\rm H}(t_{\rm rec})$ at an angle $\Delta\theta \approx 2^\circ$, as schematically shown in Fig. 17. By causality, at recombination there should be no perturbations of larger wavelengths, as any perturbation can be generated within the causal light cone only. In other words, CMB temperature must be isotropic when averaged over angular scales exceeding 2°; there should be no cold or warm spots of angular size larger than 2°. Now, CMB provides us with the photographic picture shown in Fig. 18. It is seen by the naked eye that

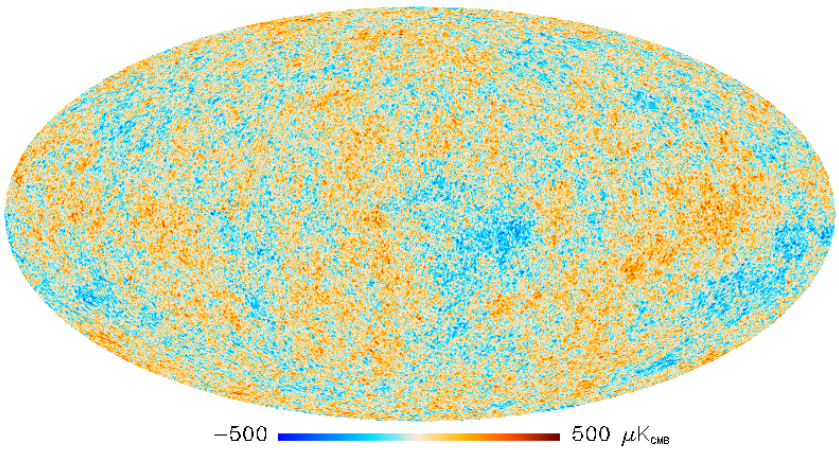


Fig. 18: CMB sky as seen by Planck

there are cold and warm regions whose angular size much exceeds 2°; in fact, there are perturbations of all angular sizes up to those comparable to the entire sky. We come to an important conclusion: *the scalar perturbations were built in at the very beginning of the hot epoch. The hot epoch was not the first, it was preceded by some other epoch, and the cosmological perturbations were generated then.*

Question. Assuming (erroneously) that there is no dark energy, and that recombination occurred deep in the matter-dominated epoch, estimate the angular scale of the horizon at recombination.

Another manifestation of the fact that the scalar perturbations were there already at the beginning of the hot epoch is the existence of peaks in the angular spectrum of CMB temperature. In general, perturbations in the baryon–photon medium before recombination are acoustic waves,

$$\delta(\mathbf{k},t) = \delta(\mathbf{k}) \mathrm{e}^{\mathrm{i}\mathbf{k}\mathbf{x}} \cos\left[\int_0^t v_{\rm s} \frac{k}{a(t')} \mathrm{d}t' + \psi_{\mathbf{k}}\right] , \tag{88}$$

where $v_{\rm s}$ is sound speed, $\delta(\mathbf{k})$ is time-independent amplitude and $\psi_{\mathbf{k}}$ is time-independent phase. This expression is valid, however, in the subhorizon regime only, i.e., at late times. The two solutions in superhorizon regime at radiation domination are

$$\delta(t) = \text{const} , \tag{89a}$$

$$\delta(t) = \frac{\text{const}}{t^{3/2}} . \tag{89b}$$

Were the perturbations generated in a causal way at radiation domination, they would be always subhorizon. In that case the solutions (89) would be irrelevant, and there would be no reason for a particular

choice of phase $\psi_{\mathbf{k}}$ in Eq. (88). One would rather expect that $\psi_{\mathbf{k}}$ is a random function of $\mathbf{k}$. This is indeed the case for specific mechanisms of the generation of density perturbations at the hot epoch [72].

On the other hand, if the perturbations existed at the very beginning of the hot epoch, they were superhorizon at sufficiently early times, and were described by the solutions (89). The consistency of the whole cosmology requires that the amplitude of perturbations was small at the beginning of the hot stage. The solution (89b) rapidly decays away, and towards the horizon entry the perturbation is in constant mode (89a). So, the initial condition for the further evolution is unique modulo amplitude $\delta(\mathbf{k})$, and hence the phase $\psi(\mathbf{k})$ is uniquely determined. For modes that enter the horizon at radiation domination this phase is equal to zero and, after entering the horizon, the modes oscillate as follows:

$$\delta(\mathbf{k},t) = \delta(\mathbf{k}) \mathrm{e}^{\mathrm{i}\mathbf{k}\mathbf{x}} \cos\left[\int_0^t v_{\mathrm{s}} \frac{k}{a(t')} \mathrm{d}t'\right] .$$

At recombination, the perturbation is

$$\delta(\mathbf{k},t_{\mathrm{r}}) = \delta(\mathbf{k}) \mathrm{e}^{\mathrm{i}\mathbf{k}\mathbf{x}} \cos(kr_{\mathrm{s}}) , \tag{90}$$

where

$$r_{\mathrm{s}} = \int_0^{t_{\mathrm{rec}}} v_{\mathrm{s}} \frac{\mathrm{d}t'}{a(t')}$$

is the comoving size of the sound horizon at recombination, while its physical size equals $a(t_{\mathrm{rec}})r_{\mathrm{s}}$. So, we see that the density perturbation of the baryon–photon plasma at recombination *oscillates as a function of wavenumber* k. The period of this oscillation is determined by r_{s}, which is a straightforwardly calculable quantity.

So, if the perturbations existed already at the beginning of the hot stage, they show the oscillatory behaviour in momentum at the recombination epoch. This translates into an oscillatory pattern of the CMB temperature angular spectrum. Omitting details, the fluctuation of the CMB temperature is partially due to the density perturbation in the baryon–photon medium at recombination. Namely, the temperature fluctuation of photons coming from the direction $\mathbf{n}$ in the sky is, roughly speaking,

$$\delta T(\mathbf{n}) \propto \delta_\gamma(\mathbf{x}_{\mathbf{n}}, \eta_{\mathrm{rec}}) + \delta T_{\mathrm{smooth}}(\mathbf{n}) ,$$

where $T_{\mathrm{smooth}}(\mathbf{n})$ corresponds to the non-oscillatory part of the CMB angular spectrum and

$$\mathbf{x}_{\mathbf{n}} = -\mathbf{n}(\eta_0 - \eta_{\mathrm{rec}}) .$$

Here $(\eta_0 - \eta_{\mathrm{rec}})$ is the coordinate distance to the sphere of photon last scattering, and $\mathbf{x}_{\mathbf{n}}$ is the coordinate of the place where the photons coming from the direction $\mathbf{n}$ scatter the last time. The quantity $T_{\mathrm{smooth}}(\mathbf{n})$ originates from the gravitational potential generated by the dark matter perturbation; dark matter has zero pressure at all times, so there are no sound waves in this component, and there are no oscillations at recombination as a function of momentum.

One expands the temperature variation on the celestial sphere in spherical harmonics:

$$\delta T(\mathbf{n}) = \sum_{lm} a_{lm} Y_{lm}(\theta,\phi).$$

The multipole number l characterizes the temperature fluctuations at the angular scale $\Delta\theta = \pi/l$. The sound waves of momentum k are seen roughly at an angle $\Delta\theta = \Delta x/(\eta_0 - \eta_{\mathrm{rec}})$, where $\Delta x = \pi/k$ is the coordinate half-wavelength. Hence, there is the correspondence

$$l \longleftrightarrow k(\eta_0 - \eta_{\mathrm{rec}}) .$$

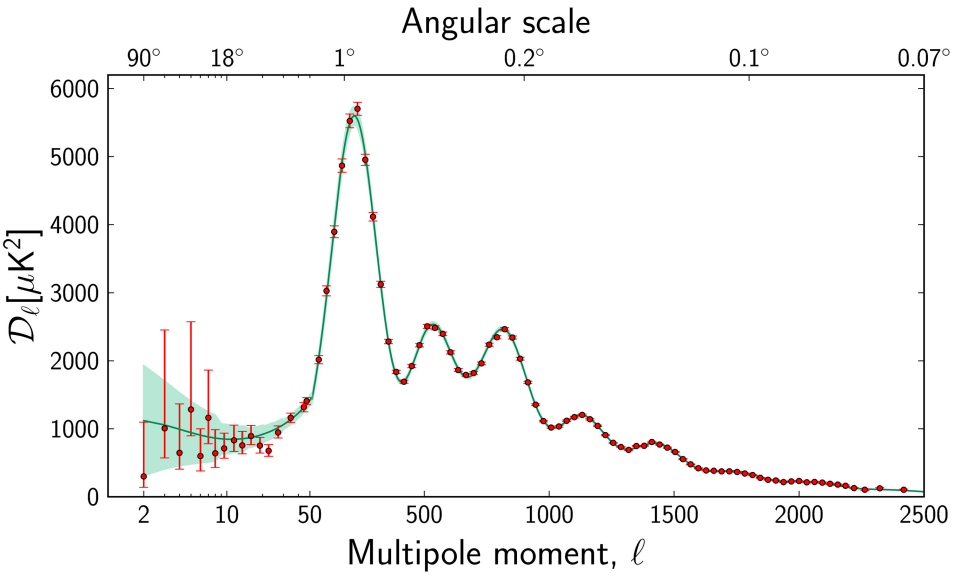


Fig. 19: The angular spectrum of the CMB temperature anisotropy [73]. The quantity on the vertical axis is D_l defined in (92). Note the unconventional scale on the horizontal axis, aimed at showing both small-l region (large angular scales) and large-l region.

Oscillations in momenta in Eq. (90) thus translate into oscillations in l, and these are indeed observed, see Fig. 19.

To understand what is shown in Fig. 19, we note that all observations today support the hypothesis that a_{lm} *are independent Gaussian random variables*. For a hypothetical ensemble of Universes like ours, the average values of products of the coefficients a_{lm} would obey

$$\langle a_{lm} a^*_{l'm'} \rangle = C_l \delta_{ll'} \delta_{mm'}. \tag{91}$$

This gives the expression for the temperature fluctuation:

$$\langle [\delta T(\mathbf{n})]^2 \rangle = \sum_l \frac{2l+1}{4\pi} C_l \approx \int \frac{\mathrm{d}l}{l} \mathcal{D}_l \,,$$

where

$$\mathcal{D}_l = \frac{l(l+1)}{2\pi} C_l \,. \tag{92}$$

Of course, one cannot measure the ensemble average (91). The definition of C_l used in experiments is

$$C_l = \frac{1}{2l+1} \sum_{m=-l}^{l} |a_{lm}|^2 \,,$$

where a_{lm} are measured quantities. Since we have only one Universe, this is generically different from the ensemble average (91): for given l, there are only $2l+1$ measurements, and the intrinsic statistical uncertainty—cosmic variance—is of order $(2l+1)^{-1/2}$. It is this uncertainty, rather than experimental error, that is shown in Fig. 19.

We conclude that the facts that the CMB angular spectrum has oscillatory behaviour and that there are sizeable temperature fluctuations at $l < 50$ (angular scale greater than the angular size of 2° of the horizon at recombination) unambiguously tell us that the density perturbations were indeed superhorizon at the hot cosmological stage. The hot epoch has to be preceded by some other epoch—the epoch of the generation of perturbations.

6.4 Primordial scalar perturbations

There are several things which we already know about the primordial density perturbations. By 'primordial' we mean the perturbations deep in the superhorizon regime at the radiation-domination epoch. As we already know, perturbations are time-independent in this regime, see Eq. (89a). They set the initial conditions for further evolution, and this evolution is well understood, at least in the linear regime. Hence, using observational data, one is able to measure the properties of primordial perturbations. Of course, since the properties we know of are established by observations, they are valid within certain error bars. Conversely, deviations from the results listed below, if observed, would be extremely interesting.

First, density perturbations are *adiabatic*. This means that there are perturbations in the energy density, but *not in composition*. More precisely, the baryon-to-entropy ratio and dark matter-to-entropy ratio are constant in space,

$$\delta\left(\frac{n_{\rm B}}{s}\right) = \text{const}\,, \qquad \delta\left(\frac{n_{\rm DM}}{s}\right) = \text{const}\,. \tag{93}$$

This is consistent with the generation of the baryon asymmetry and dark matter at the hot cosmological epoch: in that case, all particles were at thermal equilibrium early at the hot epoch, the temperature completely characterized the whole cosmic medium at that time and as long as physics behind the baryon asymmetry and dark matter generation is the same everywhere in the Universe, the baryon and dark matter abundances (relative to the entropy density) are necessarily the same everywhere. In principle, there may exist *entropy* (another term is isocurvature) perturbations, such that at the early hot epoch energy density (dominated by relativistic matter) was homogeneous, while the composition was not. This would give initial conditions for the evolution of density perturbations, which would be entirely different from those characteristic of the adiabatic perturbations. As a result, the angular spectrum of the CMB temperature anisotropy would be entirely different. No admixture of the entropy perturbations has been detected so far, but it is worth emphasizing that even a small admixture will show that many popular mechanisms for generating dark matter and/or baryon asymmetry have nothing to do with reality. One will have to think, instead, that the baryon asymmetry and/or dark matter were generated before the beginning of the hot stage. A notable example is the axion misalignment mechanism discussed in Section 4.4: in a latent sense, the axion dark matter exists from the very beginning in that case, and perturbations in the axion field $\delta\theta_0(\mathbf{x})$ (which may be generated together with the adiabatic perturbations) would show up as entropy perturbations in dark matter.

Second, the primordial density perturbations are *Gaussian random fields*. Gaussianity means that the three-point and all odd correlation functions vanish, while the four-point function and all higher order even correlation functions are expressed through the two-point function via Wick's theorem:

$$\begin{aligned} \langle\delta(\mathbf{k}_1)\delta(\mathbf{k}_2)\delta(\mathbf{k}_3)\rangle &= 0, \\ \langle\delta(\mathbf{k}_1)\delta(\mathbf{k}_2)\delta(\mathbf{k}_3)\delta(\mathbf{k}_4)\rangle &= \langle\delta(\mathbf{k}_1)\delta(\mathbf{k}_2)\rangle\cdot\langle\delta(\mathbf{k}_3)\delta(\mathbf{k}_4)\rangle \\ &\quad + \text{permutations of momenta}\,. \end{aligned}$$

We note that this property is characteristic of *vacuum fluctuations of non-interacting (linear) quantum fields.* Hence, it is quite likely that the density perturbations originate from the enhanced vacuum fluctuations of non-interacting or weakly interacting quantum field(s). The free quantum field has the general form

$$\phi(\mathbf{x},t) = \int \mathrm{d}^3k \mathrm{e}^{-\mathrm{i}\mathbf{k}\mathbf{x}}\left(f^{(+)}_{\mathbf{k}}(t)a^\dagger_{\mathbf{k}} + \mathrm{e}^{\mathrm{i}\mathbf{k}\mathbf{x}}f^{(-)}_{\mathbf{k}}(t)a_{\mathbf{k}}\right)\,,$$

where $a^\dagger_{\mathbf{k}}$ and $a_{\mathbf{k}}$ are creation and annihilation operators. For the field in Minkowski space–time, one has $f^{(\pm)}_{\mathbf{k}}(t) = \mathrm{e}^{\pm\mathrm{i}\omega_k t}$, while enhancement, e.g., due to the evolution in time-dependent background, means that $f^{(\pm)}_{\mathbf{k}}$ are large. But, in any case, Wick's theorem is valid, provided that the state of the system is vacuum, $a_{\mathbf{k}}|0\rangle = 0$.

We note in passing that *non-Gaussianity* is an important topic of current research. It would show up as a deviation from Wick's theorem. As an example, the three-point function (bispectrum) may be non-vanishing,

$$\langle\delta(\mathbf{k}_1)\delta(\mathbf{k}_2)\delta(\mathbf{k}_3)\rangle = \delta(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3)\, G(k_i^2;\ \mathbf{k}_1\mathbf{k}_2;\ \mathbf{k}_1\mathbf{k}_3) \neq 0\ .$$

The functional dependence of $G(k_i^2;\ \mathbf{k}_1\mathbf{k}_2;\ \mathbf{k}_1\mathbf{k}_3)$ on its arguments is different in different models of generation of primordial perturbations, so this shape is a potential discriminator. In some models the bispectrum vanishes, e.g., due to symmetries. In that case the trispectrum (connected four-point function) may be measurable instead. It is worth emphasizing that non-Gaussianity has not been detected yet.

Another important property is that the primordial power spectrum of density perturbations *is nearly, but not exactly, flat*. For a homogeneous and anisotropic Gaussian random field, the power spectrum completely determines its only characteristic, the two-point function. A convenient definition is

$$\langle\delta(\mathbf{k})\delta(\mathbf{k}')\rangle = \frac{1}{4\pi k^3}\mathcal{P}(k)\delta(\mathbf{k}+\mathbf{k}')\ . \tag{94}$$

The power spectrum $\mathcal{P}(k)$ defined in this way determines the fluctuation in a logarithmic interval of momenta,

$$\langle\delta^2(\mathbf{x})\rangle = \int_0^\infty \frac{\mathrm{d}k}{k}\,\mathcal{P}(k)\ .$$

By definition, the flat spectrum is such that $\mathcal{P}$ is independent of k. In this case all spatial scales are alike; no scale is enhanced with respect to another. It is worth noting that the flat spectrum was conjectured by Harrison [74], Zeldovich [75] and Peebles and Yu [76] at the beginning of the 1970s, long before realistic mechanisms of the generation of density perturbations have been proposed.

In view of the approximate flatness, a natural parametrization is

$$\mathcal{P}(k) = A_\mathrm{s}\left(\frac{k}{k_*}\right)^{n_\mathrm{s}-1}\ , \tag{95}$$

where A_s is the amplitude, $n_\mathrm{s}-1$ is the tilt and k_* is a fiducial momentum, chosen at one's convenience. The flat spectrum in this parametrization has $n_\mathrm{s}=1$. This is inconsistent with the cosmological data, which give [21]

$$n_\mathrm{s} = 0.968 \pm 0.06\ .$$

This quantifies what we mean by a nearly, but not exactly flat, power spectrum.

6.5 Inflation or not?

The pre-hot epoch must be long in terms of the time variable η introduced in Eq. (87). What we would like to have is that a large part of the Universe (e.g., the entire visible part) be causally connected towards the end of that epoch, see Fig. 20. A long duration in η does not necessarily mean a long duration in physical time t; in fact, the physical duration of the pre-hot epoch may be tiny.

An excellent hypothesis on the pre-hot stage is inflation, the epoch of nearly exponential expansion [77–82],

$$a(t) = \mathrm{e}^{\int H\mathrm{d}t}\ , \qquad H \approx \text{const}\ .$$

Inflation makes the whole visible Universe, and likely a much greater region of space, causally connected at very early times. The horizon size at inflation is at least

$$l_\mathrm{H}(t) = a(t)\int_{t_\mathrm{i}}^{t}\frac{\mathrm{d}t'}{a(t')} = H^{-1}\mathrm{e}^{H(t-t_\mathrm{i})}\ ,$$

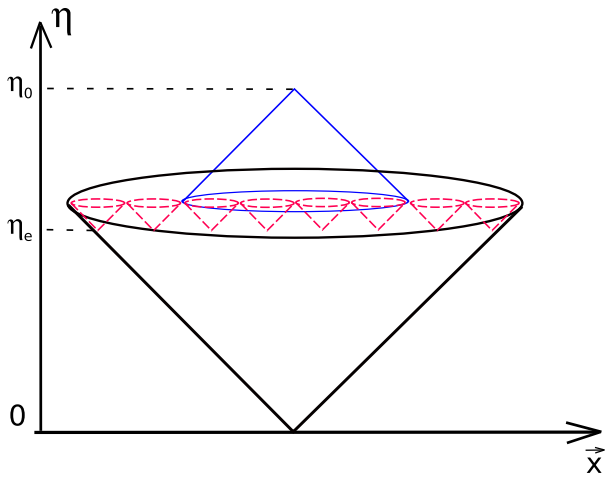


Fig. 20: Causal structure of space–time in the real Universe

where $t_{\rm i}$ is the time inflation begins, and we set $H = \text{const}$ for illustrational purposes. This size is huge for $t - t_{\rm i} \gg H^{-1}$, as desired.

Question. Assuming that at inflation $H \ll M_{\rm Pl}$, show that if the duration of inflation Δt is larger than $100H^{-1}$, the whole visible Universe is causally connected by the end of inflation. What is $100H^{-1}$ in seconds for $H = 10^{15}$ GeV? Using the time variable η, show that the causal structure of space–time in inflationary theory with $\Delta t > 100H^{-1}$ is the one shown in Fig. 20.

From the viewpoint of perturbations, the physical momentum $q(t) = k/a(t)$ decreases (gets red shifted) at inflation, while the Hubble parameter stays almost constant. So, every mode is first subhorizon ($q(t) \gg H(t)$) and later superhorizon ($q(t) \ll H(t)$). This situation is opposite to what happens at radiation and matter domination, see Fig. 21; this is precisely the prerequisite for generating the density perturbations. In fact, inflation does generate primordial density perturbations [83–87] whose properties are consistent with everything we know about them. Indeed, at the inflationary epoch, fluctuations of all

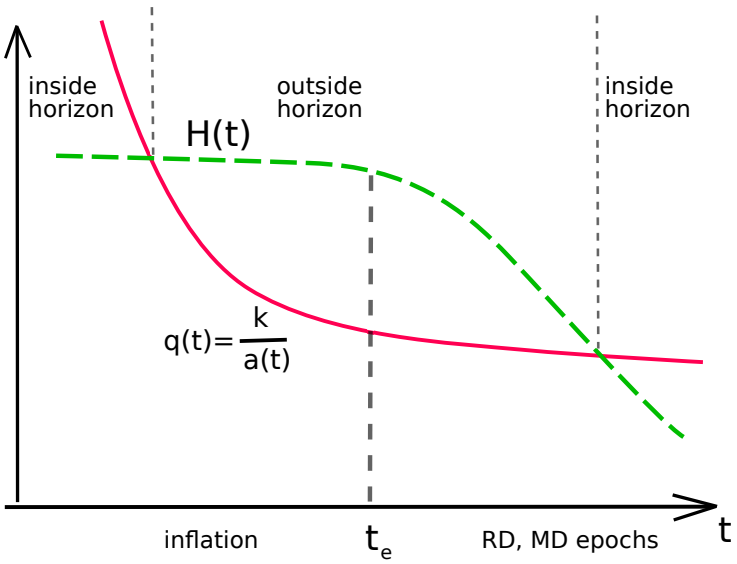


Fig. 21: Physical momentum and Hubble parameter at inflation and later: $t_{\rm e}$ is the time of the inflation end

light fields get enhanced greatly due to the fast expansion of the Universe. This is true, in particular, for the field that dominates the energy density at inflation, called an inflaton. Enhanced vacuum fluctuations of the inflaton are nothing but perturbations in the energy density at the inflationary epoch, which are reprocessed into perturbations in the hot medium after the end of inflation. The inflaton field is very weakly coupled, so the non-Gaussianity in the primordial scalar perturbations is very small [88]. In fact, it is so small that its detection is problematic even in the distant future.

The approximate flatness of the primordial power spectrum in inflationary theory is explained by the symmetry of the de Sitter space–time, which is the space–time of constant Hubble rate,

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \mathrm{e}^{2Ht}\mathrm{d}\mathbf{x}^2 \,, \qquad H = \mathrm{const} \,.$$

This metric is invariant under spatial dilatations supplemented by time translations,

$$\mathbf{x} \to \lambda\mathbf{x} \,, \quad t \to t - \frac{1}{2H}\log\lambda \,.$$

Therefore, all spatial scales are alike, which is also a defining property of the flat power spectrum. At inflation, H is almost constant in time and the de Sitter symmetry is an approximate symmetry. For this reason, inflation automatically generates a nearly flat power spectrum.

The distinguishing property of inflation is *the generation of tensor modes (primordial gravity waves)* of sizeable amplitude and nearly flat power spectrum. The gravity waves are thus a smoking gun for inflation. The reason for their generation at inflation is that the exponential expansion of the Universe enhances vacuum fluctuations of all fields, including the gravitational field itself. Particularly interesting are gravity waves whose present wavelengths are huge, 100 Mpc and larger. Many inflationary models predict their amplitudes to be very large, of order 10^{-6} or so. Shorter gravity waves are generated too, but their amplitudes decay after horizon entry at radiation domination, and today they have much smaller amplitudes making them inaccessible to gravity wave detectors like LIGO or VIRGO, pulsar timing arrays etc. A conventional characteristic of the amplitude of primordial gravity waves is the tensor-to-scalar ratio

$$r = \frac{\mathcal{P}_\mathrm{T}}{\mathcal{P}} \,,$$

where $\mathcal{P}$ is the scalar power spectrum defined in Eq. (94) and $\mathcal{P}_\mathrm{T}$ is the tensor power spectrum defined in a similar way, but for transverse traceless metric perturbations h_{ij}. The result of the search for effects of the tensor modes on CMB temperature anisotropy is shown in Fig. 22. This search has already ruled out some of the popular inflationary models.

All the above referred to the simplest, single-field inflationary models. In models with more than one relevant field, the situation may be different. In particular, sizeable non-Gaussianity may be generated, while the amplitude of tensor perturbations may be very low. So, it would be rather difficult to rule out the inflationary scenario as a whole.

Inflation is not the only hypothesis proposed so far. One option is the bouncing Universe scenario, which assumes that the cosmological evolution begins from contraction, then the contracting stage terminates at some moment of time (bounce) and is followed by expansion. A version is the cycling Universe scenario with many cycles of contraction–bounce–expansion. See reviews by Lehners and Brandenberger in Ref. [16–20]. Another scenario is that the Universe starts out from a nearly flat and static state with nearly vanishing energy density. Then the energy density increases and, according to the Friedmann equation, the expansion speeds up. This goes under the name of the Genesis scenario [90]. Theoretical realizations of these scenarios are more difficult than inflation, but they are not impossible, as became clear recently.

The generation of the density perturbations is less automatic in scenarios alternative to inflation. Similarly to inflationary theory, the flatness of the scalar power spectrum is likely to be due to some symmetry. One candidate symmetry is conformal invariance [91–94]. The point is that the conformal

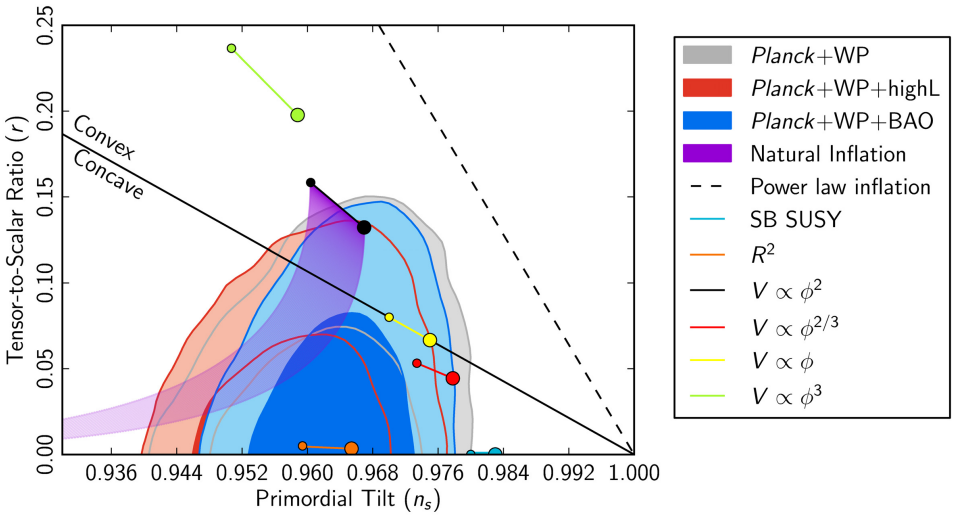


Fig. 22: Allowed regions (at 68% and 95% confidence levels) in the plane (n_s, r), where n_s is the scalar spectral index and r is the tensor-to-scalar ratio [89]. The right lower corner (the point $(1.0, 0.0)$) is the Harrison–Zeldovich point (flat scalar spectrum, no tensor modes). Intervals show predictions of popular inflationary models.

group includes dilatations, $x^\mu \to \lambda x^\mu$. This property indicates that the theory possesses no scale and has a good chance for producing the flat spectrum. A model building along this direction has begun rather recently [92–94].

6.6 Hunt continues

Until now, only very basic facts about the primordial cosmological perturbations have been observationally established. Even though very suggestive, these facts by themselves are not sufficient to unambiguously figure out what was the Universe at the pre-hot epoch of its evolution. New properties of cosmological perturbations will hopefully be discovered in the future and shed more light on this pre-hot epoch. Let us discuss some of the potential observables.

6.6.1 Tensor perturbations = relic gravity waves

As we discussed, primordial tensor perturbations are predicted by many inflationary models. On the other hand, there seems to be no way of generating a nearly flat tensor power spectrum in alternatives to inflation. In fact, most, if not all, alternative scenarios predict unobservably small tensor amplitudes. This is why we said that tensor perturbations are a smoking gun for inflation. Until recently, the most sensitive probe of the tensor perturbations has been the CMB temperature anisotropy [95–98]. However, the most promising tool is the CMB polarization. The point is that a certain class of polarization patterns (called B-mode) is generated by tensor perturbations, while scalar perturbations are unable to create it [99, 100]. Hence, dedicated experiments aiming at measuring the CMB polarization may well discover the tensor perturbations, i.e., relic gravity waves. Needless to say, this would be a profound discovery. To avoid confusion, let us note that the CMB polarization has been already observed, but it belongs to another class of patterns (so-called E-mode) and is consistent with the existence of the scalar perturbations only. The original claim of the BICEP-2 experiment [101] to detect the B-mode generated by primordial tensor perturbations was turned down [102]: the B-mode is there, but it is due to dust in our Galaxy.

6.6.2 Non-Gaussianity.

As we pointed out already, non-Gaussianity of density perturbations is very small in the simplest inflationary models. Hence, its discovery will signal that either inflation and inflationary generation of density perturbations occurred in a rather complicated way, or an alternative scenario was realized. Once the non-Gaussianity is discovered, and its shape is revealed even with modest accuracy, many concrete models will be ruled out, while at most a few will get strong support.

6.6.3 Statistical anisotropy.

In principle, the power spectrum of density perturbations may depend on the direction of momentum, e.g.,

$$\mathcal{P}(\mathbf{k}) = \mathcal{P}_0(k)\left(1 + w_{ij}(k)\frac{k_i k_j}{k^2} + \cdots\right),$$

where w_{ij} is a fundamental tensor in our part of the Universe (odd powers of k_i would contradict commutativity of the Gaussian random field $\delta(\mathbf{k})$, see Eq. (94)). Such a dependence would definitely imply that the Universe was anisotropic at the pre-hot stage, when the primordial perturbations were generated. This statistical anisotropy is rather hard to obtain in inflationary models, though it is possible in inflation with strong vector fields [103–105]. On the other hand, statistical anisotropy is natural in some other scenarios, including conformal models [106, 107]. The statistical anisotropy would show up in correlators [108, 109]

$$\langle a_{lm} a_{l'm'} \rangle \qquad \text{with } l' \neq l \text{ and/or } m' \neq m.$$

At the moment, the constraints [110, 111] on statistical anisotropy obtained by analysing the CMB data are getting into the region which is interesting from the viewpoint of some (though not many) models of the pre-hot epoch.

6.6.4 Admixture of entropy perturbations.

As we explained above, even a small admixture of entropy perturbations would force us to abandon the most popular scenarios of the generation of baryon asymmetry and/or dark matter, which assumed that it happened at the hot epoch. Once the dark matter entropy mode is discovered, the WIMP dark matter would no longer be well motivated, while other, very weakly interacting dark matter candidates, like axions or superheavy relics, would be preferred. This would redirect the experimental search for dark matter.

7 Conclusion

It is by now commonplace that the two fields studying together the most fundamental properties of matter and the Universe—particle physics and cosmology—are tightly interrelated. The present situations in these fields have much in common too. On the particle-physics side, the Standard Model has been completed by the expected discovery of the Higgs boson. On the other hand, relatively recently a fairly unexpected discovery of neutrino oscillations was made, which revolutionized our view on particles and their interactions. There are grounds to hope for even more profound discoveries, notably by the LHC experiments. While in the past there were definite predictions of the Standard Model, which eventually were confirmed, there are numerous hypotheses concerning new physics, none of which is undoubtedly plausible. On the cosmology side, the Standard Model of cosmology, ΛCDM, has been shaped, again not without an unexpected and revolutionary discovery, in this case of the accelerated expansion of the Universe. We hope for further profound discoveries in cosmology too. It may well be that we will soon learn which is the dark matter particle; again, there is an entire zoo of candidates, several of which are serious competitors. The discoveries of new properties of cosmological perturbations will hopefully reveal the nature of the pre-hot epoch. There is a clear best guess, inflation, but it is not excluded that future observational data will point towards something else.

Neither in particle physics nor in cosmology are new discoveries guaranteed, however. Nature may hide its secrets. Whether or not it does is the biggest open issue in fundamental physics.

Acknowledgement

This work is supported by the Russian Science Foundation grant 14-22-00161.

References

[1] S. Dodelson, *Modern Cosmology* (Academic Press, Amsterdam, 2003).

[2] V. Mukhanov, *Physical Foundations of Cosmology* (Cambridge University Press, Cambridge, 2005). http://dx.doi.org/10.1017/CBO9780511790553

[3] S. Weinberg, *Cosmology* (Oxford University Press, Oxford, 2008).

[4] A.R. Liddle and D.H. Lyth, *The Primordial Density Perturbation: Cosmology, Inflation and the Origin of Structure* (Cambridge University Press, Cambridge, 2009). http://dx.doi.org/10.1017/CBO9780511819209

[5] D.S. Gorbunov and V.A. Rubakov, *Introduction to the Theory of the Early Universe: Hot Big Bang Theory* (World Scientific, Hackensack, NJ, 2011). http://dx.doi.org/10.1142/7874

[6] D.S. Gorbunov and V.A. Rubakov, *Introduction to the Theory of the Early Universe: Cosmological Perturbations and Inflationary Theory* (World Scientific, Hackensack, NJ, 2011). http://dx.doi.org/10.1142/7874

[7] K.A. Olive, arXiv:astro-ph/0301505.

[8] G. Bertone, D. Hooper and J. Silk, *Phys. Rep.* **405**(5–6) (2005) 279. http://dx.doi.org/10.1016/j.physrep.2004.08.031

[9] A. Boyarsky, O. Ruchayskiy and M. Shaposhnikov, *Annu. Rev. Nucl. Part. Sci.* **59** (2009) 191. http://dx.doi.org/10.1146/annurev.nucl.010909.083654

[10] M. Kawasaki and K. Nakayama, *Annu. Rev. Nucl. Part. Sci.* **63** (2013) 69. http://dx.doi.org/10.1146/annurev-nucl-102212-170536

[11] V.A. Rubakov and M.E. Shaposhnikov, *Usp. Fiz. Nauk* **166** (1996) 493 [Engl. Trans. *Phys. Usp.* **39** (1996) 461]. http://dx.doi.org/10.3367/UFNr.0166.199605d.0493

[12] M. Trodden, *Rev. Mod. Phys.* **71**(5) (1999) 1463. http://dx.doi.org/10.1103/RevModPhys.71.1463

[13] T. Konstandin, *Usp. Fiz. Nauk* **183** (2013) 785 [Engl. trans. *Phys. Usp.* **56** (2013) 747]. http://dx.doi.org/10.3367/UFNr.0183.201308a.0785

[14] W. Buchmuller, R.D. Peccei and T. Yanagida, *Annu. Rev. Nucl. Part. Sci.* **55** (2005) 311. http://dx.doi.org/10.1146/annurev.nucl.55.090704.151558

[15] S. Davidson, E. Nardi and Y. Nir, *Phys. Rep.* **466**(4–5) (2008) 105. http://dx.doi.org/10.1016/j.physrep.2008.06.002

[16] D.H. Lyth and A. Riotto, *Phys. Rep.* **314**(1–2) (1999) 16. http://dx.doi.org/10.1016/S0370-1573(98)00128-8

[17] B.A. Bassett, S. Tsujikawa and D. Wands, *Rev. Mod. Phys.* **78**(2) (2006) 537. http://dx.doi.org/10.1103/RevModPhys.78.537

[18] J.L. Lehners, *Phys. Rep.* **465**(6) (2008) 223-263. http://dx.doi.org/10.1016/j.physrep.2008.06.001

[19] A. Mazumdar and J. Rocher, *Phys. Rep.* **497**(4–5) (2011) 85. http://dx.doi.org/10.1016/j.physrep.2010.08.001

[20] R.H. Brandenberger, *Lect. Notes Phys.* **863** (2013) 333. http://dx.doi.org/10.1007/978-3-642-33036-0_12

[21] P.A.R. Ade *et al.* (Planck Collaboration), arXiv:1502.01589 [astro-ph.CO].

[22] P.A.R. Ade *et al.* (Planck Collaboration), *Astron. Astrophys.* **571** (2014) A16. http://dx.doi.org/10.1051/0004-6361/201321591

[23] E. Gawiser and J. Silk, *Phys. Rep.* **333** (2000) 245. http://dx.doi.org/10.1016/S0370-1573(00)00025-9

[24] K. Hagiwara *et al.* (Particle Data Group Collaboration), *Phys. Rev.* D **66**(1) (2002) 010001. http://dx.doi.org/10.1103/PhysRevD.66.010001

[25] K.A. Olive *et al.* (Particle Data Group Collaboration), *Chin. Phys.* C **38**(9) (2014) 090001. http://dx.doi.org/10.1088/1674-1137/38/9/090001

[26] E. Giusarma *et al.*, *Phys. Rev.* D **90**(4) (2014) 043507. http://dx.doi.org/10.1103/PhysRevD.90.043507

[27] N.G. Busca *et al.*, *Astron. Astrophys.* **552** (2013) A96. http://dx.doi.org/10.1051/0004-6361/201220724

[28] V.A. Rubakov, *Phys. Rev.* D **61**(6) (2000) 061501. http://dx.doi.org/10.1103/PhysRevD.61.061501

[29] P.J. Steinhardt and N. Turok, *Science* **312**(5777) (2006) 1180. http://dx.doi.org/10.1126/science.1126231

[30] S. Weinberg, *Phys. Rev. Lett.* **59**(22) (1987) 2607. http://dx.doi.org/10.1103/PhysRevLett.59.2607

[31] A.D. Linde, Inflation and quantum cosmology, in *Three Hundred Years of Gravitation*, Eds S.W. Hawking and W. Israel (Cambridge University Press, Cambridge, 1987), p. 604.

[32] J.P. Kneib *et al.*, *Astrophys. J.* **598**(2) (2003) 804. http://dx.doi.org/10.1086/378633

[33] D. Clowe *et al.*, *Astrophys. J.* **648**(2) (2006) L109. http://dx.doi.org/10.1086/508162

[34] K.G. Begeman, A.H. Broeils and R.H. Sanders, *Mon. Not. R. Astron. Soc.* **249**(3) (1991) 523. http://dx.doi.org/10.1093/mnras/249.3.523

[35] T. Han, Z. Liu and A. Natarajan, *J. High Energy Phys.* **1311** (2013) 008. http://dx.doi.org/10.1007/JHEP11(2013)008

[36] A.D. Avrorin *et al.* (Baikal Collaboration), *Astropart. Phys.* **62** (2014) 12. http://dx.doi.org/10.1016/j.astropartphys.2014.07.006 http://dx.doi.org/10.1103/PhysRevLett.111.171101

[37] L. Bergstrom *et al.*, *Phys. Rev. Lett.* **111**(17) (2013) 171101. http://dx.doi.org/10.1103/physrevlett.111.171101

[38] V. Khachatryan *et al.* (CMS Collaboration), arXiv:1408.3583 [hep-ex]. http://dx.doi.org/10.1140/epjc/s10052-015-3451-4

[39] G. 't Hooft, *Phys. Rev. Lett.* **37**(1) (1976) 8. http://dx.doi.org/10.1103/PhysRevLett.37.8

[40] C.G. Callan, R.F. Dashen and D.J. Gross, *Phys. Lett.* B **63**(3) (1976) 334. http://dx.doi.org/10.1016/0370-2693(76)90277-X

[41] R. Jackiw and C. Rebbi, *Phys. Rev. Lett.* **37**(3) (1976) 172. http://dx.doi.org/10.1103/PhysRevLett.37.172

[42] J.E. Kim and G. Carosi, *Rev. Mod. Phys.* **82**(1) (2010) 557. http://dx.doi.org/10.1103/RevModPhys.82.557

[43] R.D. Peccei and H.R. Quinn, *Phys. Rev. Lett.* **38**(25) (1977) 1440. http://dx.doi.org/10.1103/PhysRevLett.38.1440

[44] S. Weinberg, *Phys. Rev. Lett.* **40**(4) (1978) 223. http://dx.doi.org/10.1103/PhysRevLett.40.223

[45] F. Wilczek, *Phys. Rev. Lett.* **40**(5) (1978) 279. http://dx.doi.org/10.1103/PhysRevLett.40.279

[46] M. Dine, W. Fischler and M. Srednicki, *Phys. Lett. B* **104**(3) (1981) 199. http://dx.doi.org/10.1016/0370-2693(81)90590-6

[47] A.R. Zhitnitsky, *Yad. Fiz.* **31**(2) (1980) 497 [Engl. trans. *Sov. J. Nucl. Phys.* **31** (1980) 260].

[48] J.E. Kim, *Phys. Rev. Lett.* **43**(2) (1979) 103. http://dx.doi.org/10.1103/PhysRevLett.43.103

[49] M.A. Shifman, A.I. Vainshtein and V.I. Zakharov, *Nucl. Phys.* B **166**(3) (1980) 493. http://dx.doi.org/10.1016/0550-3213(80)90209-6

[50] A. Vilenkin and A.E. Everett, *Phys. Rev. Lett.* **48**(26) (1982) 1867. http://dx.doi.org/10.1103/PhysRevLett.48.1867

[51] R.A. Battye and E.P.S. Shellard, arXiv:astro-ph/9909231.

[52] J. Preskill, M.B. Wise and F. Wilczek, *Phys. Lett.* B **120**(1–3) (1983) 127. http://dx.doi.org/10.1016/0370-2693(83)90637-8

[53] L.F. Abbott and P. Sikivie, *Phys. Lett.* B **120**(1–3) (1983) 133. http://dx.doi.org/10.1016/0370-2693(83)90638-X

[54] M. Dine and W. Fischler, *Phys. Lett.* B **120**(1–3) (1983) 137. http://dx.doi.org/10.1016/0370-2693(83)90639-1

[55] A. Ringwald, *J. Phys. Conf. Ser.* **485** (2014) 012013. http://dx.doi.org/10.1088/1742-6596/485/1/012013

[56] D. Gorbunov, A. Khmelnitsky and V. Rubakov, *J. Cosmol. Astropart. Phys.* **0810** (2008) 041. http://dx.doi.org/10.1088/1475-7516/2008/10/041

[57] M. Laine and M. Shaposhnikov, *J. Cosmol. Astropart. Phys.* **0806** (2008) 031. http://dx.doi.org/10.1088/1475-7516/2008/06/031

[58] X.-D. Shi and G.M. Fuller, *Phys. Rev. Lett.* **82**(14) (1999) 2832. http://dx.doi.org/10.1103/PhysRevLett.82.2832

[59] D. Gorbunov, A. Khmelnitsky and V. Rubakov, *J. High Energy Phys.* **0812** (2008) 055. http://dx.doi.org/10.1088/1126-6708/2008/12/055

[60] A. de Gouvea, T. Moroi and H. Murayama, *Phys. Rev.* D **56**(2) (1997) 1281. http://dx.doi.org/10.1103/PhysRevD.56.1281

[61] A.D. Sakharov, *Pisma Zh. Eksp. Teor. Fiz.* **5** (1967) 32 [Engl. trans. *JETP Lett.* **5** (1967) 24]. http://dx.doi.org/10.1070/PU1991v034n05ABEH002497

[62] V.A. Kuzmin, *Pisma Zh. Eksp. Teor. Fiz.* **12** (1970) 335.

[63] I. Affleck and M. Dine, *Nucl. Phys.* B **249**(2) (1985) 361. http://dx.doi.org/10.1016/0550-3213(85)90021-5

[64] F.R. Klinkhamer and N.S. Manton, *Phys. Rev.* D **30**(10) (1984) 2212. http://dx.doi.org/10.1103/PhysRevD.30.2212

[65] V.A. Kuzmin, V.A. Rubakov and M.E. Shaposhnikov, *Phys. Lett.* B **155**(1–2) (1985) 36. http://dx.doi.org/10.1016/0370-2693(85)91028-7

[66] V.A. Rubakov, *Classical Theory of Gauge Fields* (Princeton University Press, Princeton, NJ, 2002).

[67] A.A. Belavin *et al.*, *Phys. Lett.* B **59**(1) (1975) 85. http://dx.doi.org/10.1016/0370-2693(75)90163-X

[68] M.S. Carena, M. Quiros and C.E.M. Wagner, *Phys. Lett.* B **380**(1–2) (1996) 81. http://dx.doi.org/10.1016/0370-2693(96)00475-3

[69] M.S. Carena, *et al.*, *Nucl. Phys.* B **503**(1–2) (1997) 387. http://dx.doi.org/10.1016/S0550-3213(97)00412-4

[70] M.S. Carena *et al.*, *Nucl. Phys.* B **650**(1–2) (2003) 24. http://dx.doi.org/10.1016/S0550-3213(02)01065-9

[71] M. Fukugita and T. Yanagida, *Phys. Lett.* B **174**(1) (1986) 45. http://dx.doi.org/10.1016/0370-2693(86)91126-3

[72] J. Urrestilla *et al.*, *J. Cosmol. Astropart. Phys.* **0807** (2008) 010. http://dx.doi.org/10.1088/1475-7516/2008/07/010

[73] P.A.R. Ade *et al.* (Planck Collaboration), *Astron. Astrophys.* **571** (2014) A1. http://dx.doi.org/10.1051/0004-6361/201321529

[74] E.R. Harrison, *Phys. Rev.* D **1**(10) (1970) 2726. http://dx.doi.org/10.1103/PhysRevD.1.2726

[75] Y.B. Zeldovich, *Mon. Not. R. Astron. Soc.* **160** (1972) 1P. http://dx.doi.org/10.1093/mnras/160.1.1P

[76] P.J.E. Peebles and J.T. Yu, *Astrophys. J.* **162** (1970) 815. http://dx.doi.org/10.1086/150713

[77] A.A. Starobinsky, *Pisma Zh. Eksp. Teor. Fiz.* **30** (1979) 719 [Engl. trans. *JETP Lett.* **30** (1979) 682].

[78] A.A. Starobinsky, *Phys. Lett.* B **91**(1) (1980) 99. http://dx.doi.org/10.1016/0370-2693(80)90670-X

[79] A.H. Guth, *Phys. Rev.* D **23**(2) (1981) 347. http://dx.doi.org/10.1103/PhysRevD.23.347

[80] A.D. Linde, *Phys. Lett.* B **108**(6) (1982) 389. http://dx.doi.org/10.1016/0370-2693(82)91219-9

[81] A. Albrecht and P.J. Steinhardt, *Phys. Rev. Lett.* **48**(17) (1982) 1220. http://dx.doi.org/10.1103/PhysRevLett.48.1220

[82] A.D. Linde, *Phys. Lett.* B **129**(3-4) (1983) 177. http://dx.doi.org/10.1016/0370-2693(83)90837-7

[83] V.F. Mukhanov and G.V. Chibisov, *Pisma Zh. Eksp. Teor. Fiz.* **33** (1981) 549 [Engl. trans. *JETP Lett.* **33** (1981) 532].

[84] S.W. Hawking, *Phys. Lett.* B **115**(4) (1982) 295. http://dx.doi.org/10.1016/0370-2693(82)90373-2

[85] A.A. Starobinsky, *Phys. Lett.* B **117**(3–4) (1982) 175. http://dx.doi.org/10.1016/0370-2693(82)90541-X

[86] A.H. Guth and S.Y. Pi, *Phys. Rev. Lett.* **49**(15) (1982) 1110. http://dx.doi.org/10.1103/PhysRevLett.49.1110

[87] J.M. Bardeen, P.J. Steinhardt and M.S. Turner, *Phys. Rev.* D **28**(4) (1983) 679. http://dx.doi.org/10.1016/0370-2693(82)91219-9

[88] J.M. Maldacena, *J. High Energy Phys.* **0305** (2003) 013. http://dx.doi.org/10.1088/1126-6708/2003/05/013

[89] P.A.R. Ade *et al.* (Planck Collaboration), *Astron. Astrophys.* **571** (2014) A22. http://dx.doi.org/10.1051/0004-6361/201321569

[90] P. Creminelli, A. Nicolis and E. Trincherini, *J. Cosmol. Astropart. Phys.* **1011** (2010) 021. http://dx.doi.org/10.1088/1475-7516/2010/11/021

[91] I. Antoniadis, P.O. Mazur and E. Mottola, *Phys. Rev. Lett.* **79**(1) (1997) 14. http://dx.doi.org/10.1103/PhysRevLett.79.14

[92] V.A. Rubakov, *J. Cosmol. Astropart. Phys.* **0909** (2009) 030. http://dx.doi.org/10.1088/1475-7516/2009/09/030

[93] P. Creminelli, A. Nicolis and E. Trincherini, *J. Cosmol. Astropart. Phys.* **1011** (2010) 021. http://dx.doi.org/10.1088/1475-7516/2010/11/021

[94] K. Hinterbichler and J. Khoury, *J. Cosmol. Astropart. Phys.* **1204** (2012) 023. http://dx.doi.org/10.1088/1475-7516/2012/04/023

[95] V.A. Rubakov, M.V. Sazhin and A.V. Veryaskin, *Phys. Lett.* B **115**(3) (1982) 189. http://dx.doi.org/10.1016/0370-2693(82)90641-4

[96] R. Fabbri and M.D. Pollock, *Phys. Lett.* B **125**(6) (1983) 445. http://dx.doi.org/10.1016/0370-2693(83)91322-9

[97] L.F. Abbott and M.B. Wise, *Nucl. Phys.* B **244**(2) (1984) 541. http://dx.doi.org/10.1016/0550-3213(84)90329-8

[98] A.A. Starobinsky, *Sov. Astron. Lett.* **11** (1985) 133.

[99] M. Kamionkowski, A. Kosowsky and A. Stebbins, *Phys. Rev. Lett.* **78**(11) (1997) 2058. http://dx.doi.org/10.1103/PhysRevLett.78.2058

[100] U. Seljak and M. Zaldarriaga, *Phys. Rev. Lett.* **78**(11) (1997) 2054. http://dx.doi.org/10.1103/PhysRevLett.78.2054

[101] P.A.R. Ade *et al.* (BICEP2 Collaboration), *Phys. Rev. Lett.* **112**(24) (2014) 241101. http://dx.doi.org/10.1103/PhysRevLett.112.241101

[102] P.A.R. Ade *et al.* (BICEP2 and Planck Collaborations), *Phys. Rev. Lett.* **114**(10) (2015) 101301. http://dx.doi.org/10.1103/PhysRevLett.114.101301

[103] M.A. Watanabe, S. Kanno and J. Soda, *Phys. Rev. Lett.* **102**(19) (2009) 191302. http://dx.doi.org/10.1103/PhysRevLett.102.191302

[104] T.R. Dulaney and M.I. Gresham, *Phys. Rev.* D **81**(10) (2010) 103532. http://dx.doi.org/10.1103/PhysRevD.81.103532

[105] A.E. Gumrukcuoglu, B. Himmetoglu and M. Peloso, *Phys. Rev.* D **81**(6) (2010) 063528. http://dx.doi.org/10.1103/PhysRevD.81.063528

[106] M. Libanov and V. Rubakov, *J. Cosmol. Astropart. Phys.* **1011** (2010) 045. http://dx.doi.org/10.1088/1475-7516/2010/11/045

[107] M. Libanov, S. Ramazanov and V. Rubakov, *J. Cosmol. Astropart. Phys.* **1106** (2011) 010. http://dx.doi.org/10.1088/1475-7516/2011/06/010

[108] L. Ackerman, S.M. Carroll and M.B. Wise, *Phys. Rev.* D **75** (2007) 083502 [Erratum *Phys. Rev.* D **80** (2009) 069901]. http://dx.doi.org/10.1103/PhysRevD.75.083502

[109] A.R. Pullen and M. Kamionkowski, *Phys. Rev.* D **76**(10) (2007) 103529. http://dx.doi.org/10.1103/PhysRevD.76.103529

[110] J. Kim and E. Komatsu, *Phys. Rev.* D **88**(10) (2013) 101301. http://dx.doi.org/10.1103/PhysRevD.88.101301

[111] G.I. Rubtsov and S.R. Ramazanov, *Phys. Rev.* D **91**(4) (2015) 043514. http://dx.doi.org/10.1103/PhysRevD.91.043514

Proceedings of the 2014 Asia–Europe–Pacific School of High-Energy Physics, Puri, India, 4–17 November 2014, edited by
M. Mulders and R. Godbole, CERN Yellow Reports: School Proceedings, Vol. 2/2017, CERN-2017-005-SP (CERN, Geneva, 2017)

Particle Physics Instrumentation

I. Wingerter-Seez
LAPP, CNRS, Paris, France, and Université Savoie Mont Blanc, Chambéry, France

Abstract
This reports summarizes the three lectures on particle physics instrumentation given during the AEPSHEP school in November 2014 at Puri-India. The lectures were intended to give an overview of the interaction of particles with matter and basic particle detection principles in the context of large detector systems like the Large Hadron Collider.

Keywords
Lectures; instrumentation; particle physics; detector; energy loss; multiple scattering.

1 Introduction

This report gives a very brief overview of the basis of particle detection and identification in the context of high-energy physics (HEP). It is organized into three main parts. A very simplified description of the interaction of particles with matter is given in Sections 2 and 3; an overview of the development of electromagnetic (EM) and hadronic showers is given in Section 4. Sections 5 and 6 are devoted to the principle of operation of gas- and solid-state detectors and calorimetry. Finally, Sections 7 and 8 give a rapid overview of the HEP detectors and a few examples.

2 Basics of particle detection

Experimental particle physics is based on high precision detectors and methods. Particle detection exploits the characteristic interaction with matter of a few well-known particles. The Standard Model of particle physics is based on 12 elementary fermion particles (six leptons, six quarks), four types of spin-1 bosons (photon, W, Z, and gluon) and the spin-0 Higgs boson. All the known particles are combinations of the elementary fermions. Most of the known particles are unstable hadrons which decay before reaching any sensitive detector. Among the few hundreds of the known particles, eight of them are the most frequently used for detection: electron, muon, photon, charged pion, charged kaon, neutral kaon, proton and neutron.

The interaction of these particles with matter constitutes the key to detection and identification. Figure 1 schematically illustrates the results of particle interactions in a typical HEP detector. The difference in mass, charge, and type of interaction constitute the means to their identification. The electron leaves a track in the tracking detector and creates an EM shower in the calorimeter. The photon may either traverse the tracking detector and only interact in the calorimeter, initiating an EM shower or it may convert to a pair $\mathrm{e}^+\mathrm{e}^-$ in the tracking detector matter. The muon, which has a mass 200 times larger than the electron, traverses the entire detector, leaving a track in the central and the muon tracking systems. Charged hadrons such as $\pi^\pm$ and $\mathrm{K}^\pm$ protons leave a track in the tracking detector and deposit their energy in the calorimeters. The neutron and the neutral kaon K^0 do not leave a track and produce a hadron shower in the calorimeter. A neutrino traverses the entire detector without interacting but its presence can be detected via energy balance.

The role of a particle detector is to detect the passage of a particle, localize its position, measure its momentum or energy, identify its nature, and measure its arrival time. Detection happens through particle energy loss in the traversed material. The detector converts this energy loss to a detectable signal which is collected and interpreted.

https://doi.org/10.23730/CYRSP-2017-002.295

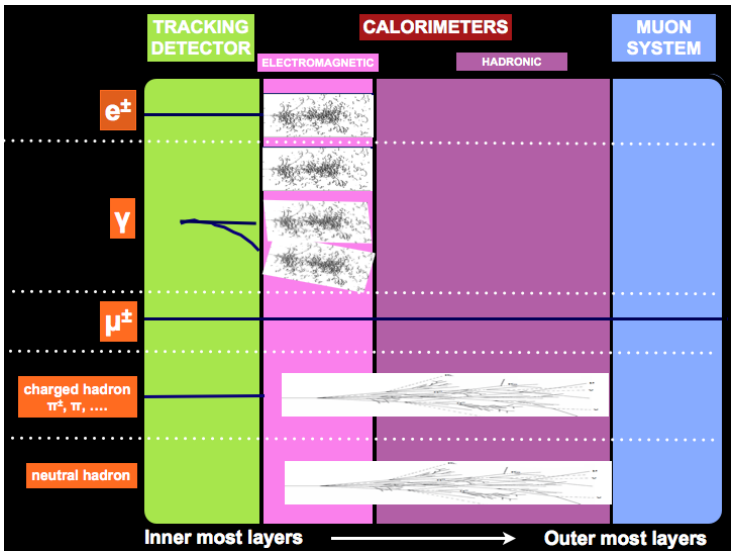


Fig. 1: Schematic representation of the passage of one electron, photon (unconverted and converted), $\mu^{\pm}$, charged hadron, and neutral hadron in a typical HEP detector built from a tracking detector, a calorimeter, and a muon detector.

3 Interaction of particles with matter

The EM interaction of charged particles with matter constitutes the essence of particle detection. Four main components of EM interaction can be identified: interaction with atomic electrons, interaction with the atomic nucleus, and two long-range collective effects, Cherenkov and transition radiations. Interactions with atomic electrons leads to ionization and excitation, and interactions with the nucleus lead to Compton scattering, bremsstrahlung, and pair production (for photons). Hadrons are sensitive to the nuclear force and therefore also obey the nuclear interaction.

This section highlights a few prominent characteristics of particle interaction with matter. For a complete treatment and a list of references, see Ref. [1].

3.1 Ionization and excitation

A heavy ($m \gg m_{\mathrm{e}}$) charged particle with $0.01 < \beta\gamma < 1000$, passing through an atom will interact via the Coulomb force with the atomic electrons and the nucleus. Because of the large mass difference ($2m_{\mathrm{p}}/m_{\mathrm{e}}$) between the atomic electrons and the nucleus, the energy transfer to the atomic electrons dominates (typically by a factor of 4000). The distribution of average energy transfer per unit distance (also called stopping power) of positively charged muons impinging on copper is presented in Fig. 2 [1].

3.1.1 Energy loss for heavy particles

The average energy transfer, for incoming heavy particles ($m \gg m_{\mathrm{e}}$) with $0.01 < \beta\gamma < 100$, is well described by the Bethe formula which is reproduced in Eq. (1) [1] (terms for this are defined in Table 1)

$$-\frac{\mathrm{d}E}{\mathrm{d}x} = Kz^2\frac{Z}{A}\frac{1}{\beta^2}\left[\frac{1}{2}\ln\frac{2m_e c^2\beta^2\gamma^2 W_{\max}}{I^2} - \beta^2 - \frac{\delta(\beta\gamma)}{2}\right] \quad \left[\,\mathrm{MeV\ g^{-1}\ cm^2}\right] . \tag{1}$$

The energy loss in a given material is first-order independent of the mass of the incoming particle: Eq. (1) and the curve of Fig. 2 can therefore be considered universal. The energy loss is proportional to the square of the incoming particle charge: for instance, a helium nucleus deposits four times more energy than a proton. From knowing the energy loss and ionization energy of a material, it is possible

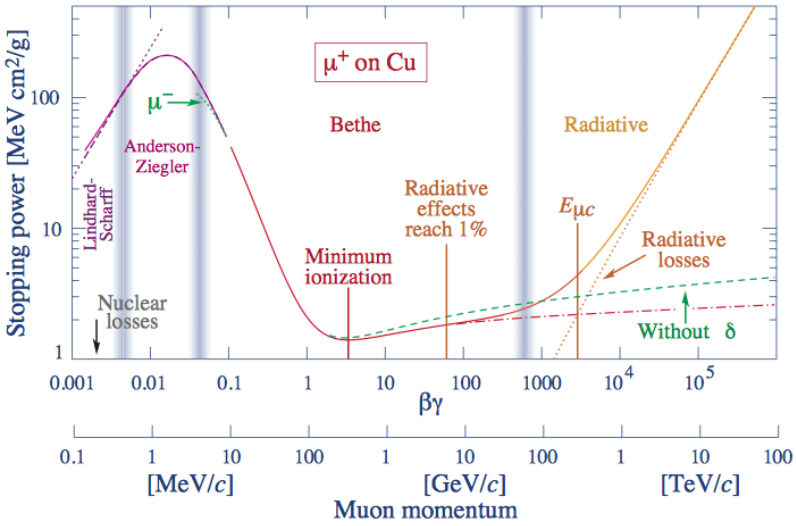


Fig. 2: Stopping power $-\langle \frac{\mathrm{d}E}{\mathrm{d}x} \rangle$ for positively charged muons in copper as a function of $\beta\gamma$ over nine orders of magnitude in momentum.

Table 1: Bethe–Bloche formula terms

$\frac{\mathrm{d}E}{\mathrm{d}x}$	Average energy loss in unit of MeV g^{-1} cm^2
$K = 4\pi N_\mathrm{A} r_\mathrm{e}^2 m_\mathrm{e} c^2 = 0.307$ MeV g^{-1} cm^2	
$W_\mathrm{max} = 2m_e c^2 \beta^2\gamma^2/(1 + 2\gamma m_\mathrm{e}/M + (m_\mathrm{e}/M)^2)$	Maximum energy transfer in a single collision
z	Charge of the incident particle
M	Mass of the incident particle
Z	Charge number of the medium
A	Atomic mass of the medium
I	Mean excitation energy of the medium
δ	Density correction (transverse extension of the electric field)
$N_\mathrm{A} = 6.022 \times 10^{23}$	Avogadro's number
$r_\mathrm{e} = e^2/4\pi\epsilon_0 m_\mathrm{e} c^2$ = 2.8 fm	Classical electron radius
m_e = 511 keV	Electron mass
$\beta = v/c$	Velocity
$\gamma = (1 - \beta^2)^{-1/2}$	Lorenz factor

to compute the number of electron–ion pairs created along the path of the traversing particle. For a muon at minimum ionization energy traversing 1 cm of copper, the number of electron–ion pairs is $\simeq 13 \times 10^6/7.7 \simeq 2 \times 10^6$ with the copper ionization energy being 7.7 eV.

The energy loss by incoming particles leads to two effects, depending on the distance to the atomic electrons. If the distance is large, the transferred energy will not be large enough for the electron to be extracted from the atom, and the atomic electron will go into an excited state and emit photons. If the distance is smaller, the transferred energy can be above the binding energy, the electron will be freed, and the atom ionized. The photons resulting from the de-excitation of the atoms and the ionization electrons and ions are used to generate signals that can be readout by the detector.

The understanding of energy loss of heavy particles (with $m \gg m_\mathrm{e}$) can be analysed in four main regimes.

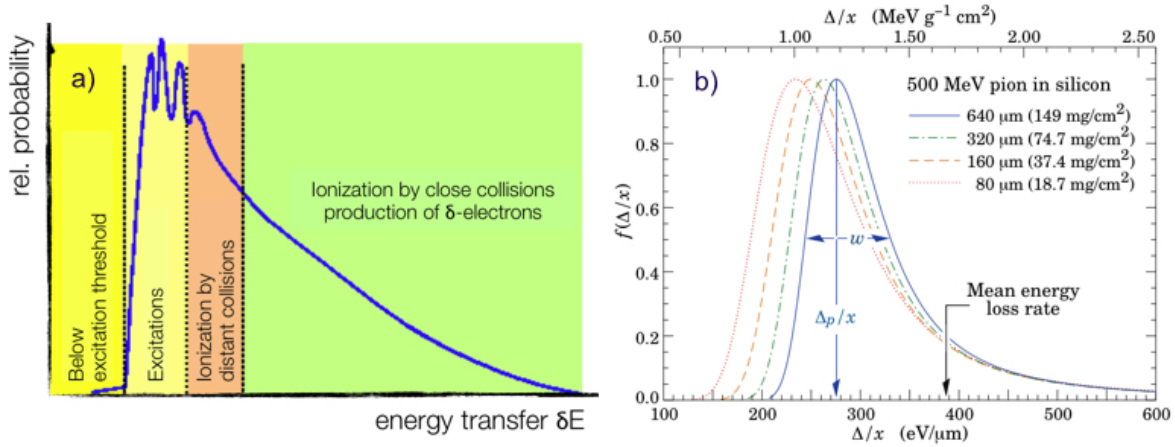


Fig. 3: (a) Schematic representation of the probability of energy transfer. (b) Energy loss probability described by Landau functions for incoming 500 MeV charged pions in various thickness of silicon.

1. The minimum ionization at $\beta\gamma \simeq 3$–4 is typical and more or less universal for materials: $-\frac{\mathrm{d}E}{\mathrm{d}x} \simeq$ 1–2 MeV g^{-1} cm^2. As an example, for copper which has a density of 8.94 g cm^{-3}, a particle at minimum ionization deposits $\simeq$13 MeV cm^{-1}.
2. For $\beta\gamma < 3$–4, the energy loss decreases as the momentum increases, with a dependency of $\simeq\beta^{-2}$ as slower particles feel the electric force of atomic electrons for a longer time.
3. For $\beta\gamma > 4$, when the particle velocity approaches the speed of light, the decrease in energy loss should reach a minimum. However, the relativistic effect induces the increase of the transverse electric field, with a dependency in $\ln\gamma$, and the interaction cross-section increases. This effect is called the relativistic rise.
4. Some further corrections to the simplified Bethe formula are necessary to account for effects from density at high γ values and for effects at very low values ($\beta\gamma < 0.1 - 1$) where the particle velocity is close to the orbital velocity of electrons.

As the energy loss is a function of the velocity, measuring the incoming particle momentum inside a magnetic field makes it possible to identify the particle by relating the momentum and the energy deposition, typically in the low momentum region where the deposited energy is large.

The Bethe formula describes the mean energy deposited by a high-mass particle in a medium. The energy deposition is a statistical process which can be described by the succession of energy loss in a material of thickness Δx, with total energy deposition $\Delta E = \sum_{n=1}^{N} \delta E_n$ following the probability density function presented in Fig. 3(a). Energy loss is well described by a Landau distribution as illustrated in Fig. 3(b).

3.1.2 Energy loss for electrons and positrons

The Bethe formula given in Eq. (1) is valid for heavy particles and needs to be modified to describe the energy loss of incoming electrons and positrons because the mass of the incoming particle and its target are the same. The formula for electrons is presented as

$$-\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{1}{2}K\frac{Z}{A}\frac{1}{\beta^2}\left[\ln\frac{m_e c^2\beta^2\gamma^2\left(m_e c^2(\gamma-1)/2\right)}{2I^2} + F(\gamma)\right] \quad \left[\text{MeV g}^{-1}\text{ cm}^2\right] . \tag{2}$$

For positrons, this formula is different again because of the difference between electrons and positrons and to account for the annihilation cross-section at low energies.

3.2 Multiple scattering, bremsstrahlung and pair production

3.2.1 Multiple scattering

The Coulomb interaction of an incoming particle with the atomic nuclei of the detector material results in the deflection of the particle, which is called multiple scattering. The statistical sum of many such small scattering angles results in a Gaussian angular distribution with a width, θ_0 given by

$$\theta_0 = \frac{13.6 \text{ MeV}}{\beta cp} z\sqrt{x/X_0}\left[1 + 0.038 \ln\left(x/X_0\right)\right] , \tag{3}$$

where x is the distance traversed and X_0 is the radiation length which is introduced in Section 3.2.2, see Eq. (6).

Multiple scattering is an intrinsic limiting factor for tracking detectors as it induces an irreducible contribution to the resolution. For example, the standard deviation for the multiple scattering of a 1 GeV incoming pion, traversing 300 μm of silicon ($X_0 = 9.4$ cm), is $\theta_0 = 0.8$ mrad, corresponding to a distance of 80 μm after 10 cm which is significantly larger than the typical resolution of a silicon strip detector.

As shown in Eq. (3), θ_0 is inversely proportional to the particle velocity and momentum, and loss of momentum resolution from multiple scattering is greater for low-energy particles. The standard deviation is also proportional to the square root of the material thickness in units of radiation length. Reducing the material thickness will lead to a reduction of the contribution the multiple scattering. Tracking devices therefore favour thin material with a small radiation length, i.e. with a low atomic number (see Eq. (6)), such as beryllium, used for beam pipes, carbon fibre and aluminum, used for support structures, and thin silicon or gas detectors.

3.2.2 Interactions of high-energy electrons with matter

The interaction of intermediate-energy electrons with matter is described in this section. For incoming electrons and positrons at higher energies, the EM interaction with the nucleus is dominant. The deflection of the charged particle by the nuclei results in acceleration and emission of EM radiation. This effect, called bremsstrahlung, plays a key role in calorimetry measurements. The bremsstrahlung process is represented in Fig. 4

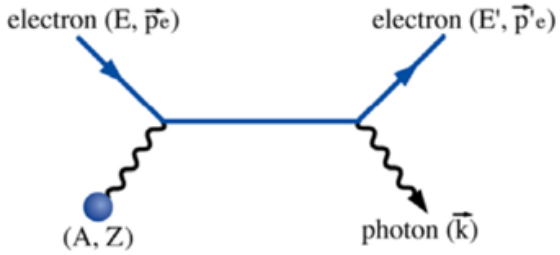


Fig. 4: The incoming electron (E, $\vec{p_e}$) interacts with the nucleus (A, Z) of the traversed matter; the emitted photon carries a momentum $\vec{k}$ and the outgoing electron (E′, $\vec{p_e}'$).

The spectrum of photons with energy k, radiated by an electron traversing a thin slab of material has the following characteristic bremsstrahlung spectrum (dominantly in $\frac{1}{k}$) expressed as a function of $y = k/E$ [9]:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}k} = 4\alpha Z(Z+1)r_e^2 \ln\left(183Z^{-1/3}\right)\left(\frac{4}{3} - \frac{3}{4}y + y^2\right) \times \frac{1}{k} , \tag{4}$$

where r_{e}^2 is the classical radius of the electron.

The term $Z(Z+1)$ of Eq. (4) reflects that the bremsstrahlung results from coupling of the initial electron to the EM field of the nucleus, augmented by a direct contribution of the atomic electrons (Z^2 replaced by $Z(Z+1)$). The logarithmic term $\ln\left(183Z^{-1/3}\right)$ shows that the atomic electrons screen the nucleus field.

For a given energy E, the average energy lost by bremsstrahlung, $\mathrm{d}E$, in a thin slab of material of thickness, $\mathrm{d}x$, is obtained by integrating over y.

$$\frac{\mathrm{d}E}{\mathrm{d}x} = 4\alpha N_{\mathrm{A}} z^2 Z^2 \left(\frac{1}{4\pi\epsilon_0}\frac{e^2}{mc^2}\right) E \ln\left(183Z^{-1/3}\right) \propto \frac{E}{m^2} \,. \tag{5}$$

As shown in Eq. (5), the energy loss is proportional to $\frac{1}{m^2}$ and is therefore mostly relevant for electrons. For a given energy, as $m_\mu \approx 200 \times m_{\mathrm{e}}$, the energy loss for a muon is $\approx 40\,000$ times smaller than for an electron. In other words, the average energy loss by bremsstrahlung dominates ionization for muons with energy above 400 GeV.

Following Eq. (5) the radiation length X_0 can be defined as the distance after which the incident electron has radiated $(1-\frac{1}{\mathrm{e}})$ = 63% of its incident energy and one can then write

$$E(x) = E_0 \mathrm{e}^{-x/X_0} \quad \text{where } x \text{ is the depth in the block of matter and}$$

$$\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{E}{X_0} \quad \text{where } X_0 = 4\alpha N_{\mathrm{A}} Z^2 r_{\mathrm{e}}^2 \ln 183Z^{-1/3} \,. \tag{6}$$

The critical energy E_{c} (or ϵ_0), at which the average energy loss by ionization equals the average energy loss by ionization, constitutes a useful quantity to describe the development of EM showers. Approximation of E_c, for gas and liquid or solid, is given as

$$E_{\mathrm{c}}^{\text{gas}} = \frac{710\ \text{MeV}}{Z+0.92} \quad \text{and} \quad E_{\mathrm{c}}^{\text{sol/liq}} = \frac{610\ \text{MeV}}{Z+1.24} \,. \tag{7}$$

Figure 5 [1] shows the fractional energy loss per radiation length as a function of electron or positron energy. Contributions from low-energy processes such as Moller and Bhabha scattering and $\mathrm{e}^+\mathrm{e}^-$ annihilation are shown.

3.2.3 Interactions of photons with matter

For high-energy photons for which $E > 2m_{\mathrm{e}}c^2$, pair creation is the dominant process as represented in Fig. 6. Similarly to the bremsstrahlung process for electrons, the pair creation results from the EM interaction between the incoming photon and the field of the atom nucleus. The pair creation cross-section is given as

$$\frac{\mathrm{d}\sigma}{\mathrm{d}x} = \frac{A}{X_0 N_{\mathrm{A}}} \times \left(1 - \frac{4}{3}x(1-x)\right) \,, \tag{8}$$

where $x = \frac{E}{k}$ is the fraction of the energy of the incoming photon carried by the produced electron. The total pair production cross-section is given as

$$\sigma_{\text{pair}} = \frac{7}{9}\frac{A}{X_0 N_{\mathrm{A}}} = \frac{7}{9} 4\alpha Z(Z+1) r_{\mathrm{e}}^2 \ln\left(183Z^{-1/3}\right) \,. \tag{9}$$

The dominant part Z^2 is due to the interaction with the nucleus; atomic electrons contribute proportionally to Z. The electron and positron are collinear as the energy of recoil of the nucleus is small ($\simeq m_{\mathrm{e}}c^2$).

In addition to pair creation, photons interact in several ways.

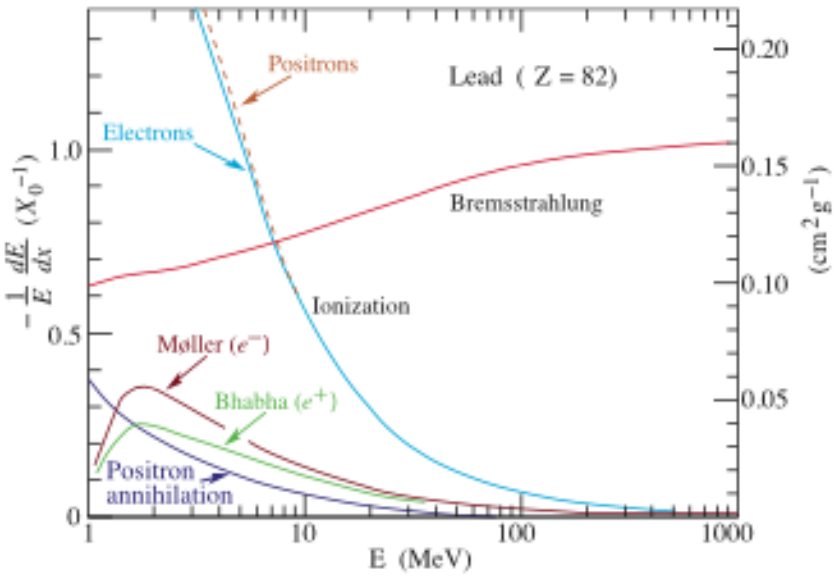


Fig. 5: Fractional energy loss per radiation length as a function of electron or positron energy in lead

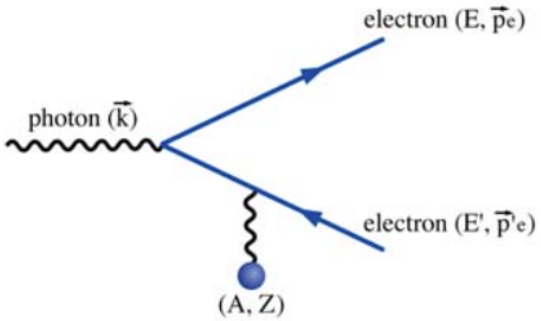


Fig. 6: The incoming photon ($\vec{k}$) interacts with the EM field of the nucleus (A, Z): an electron ($\mathrm{E}, \vec{p_\mathrm{e}}$), positron ($\mathrm{E}', \vec{p_\mathrm{e}}'$) pair is created.

- **Photo electric effect**: for low-energy photons, the atomic electrons are not free; therefore, the cross-section presents discontinuities whenever the photon energy crosses the electron binding energy. This process is strongly Z dependent ($\frac{Z^5}{E^{3.5}}$) and is dominant at very low energies.
- **Compton scattering**: scattering of the incoming on one atomic electron. The cross-section varies like $Z\frac{\ln E}{E}$.

Photo-electrons and scattered electrons are emitted isotropically whereas electrons produced by pair creation are emitted in the direction of the incoming photon. Figure 7 from Ref. [1] presents the total cross-section of photons impinging on carbon (left) and on lead (right).

High-energy electrons, positrons, and photons ($E > 100$ MeV) impinging on a block of material interact dominantly with the atom nucleus via bremsstrahlung and pair-creation processes. These two processes dominate the development of EM showers.

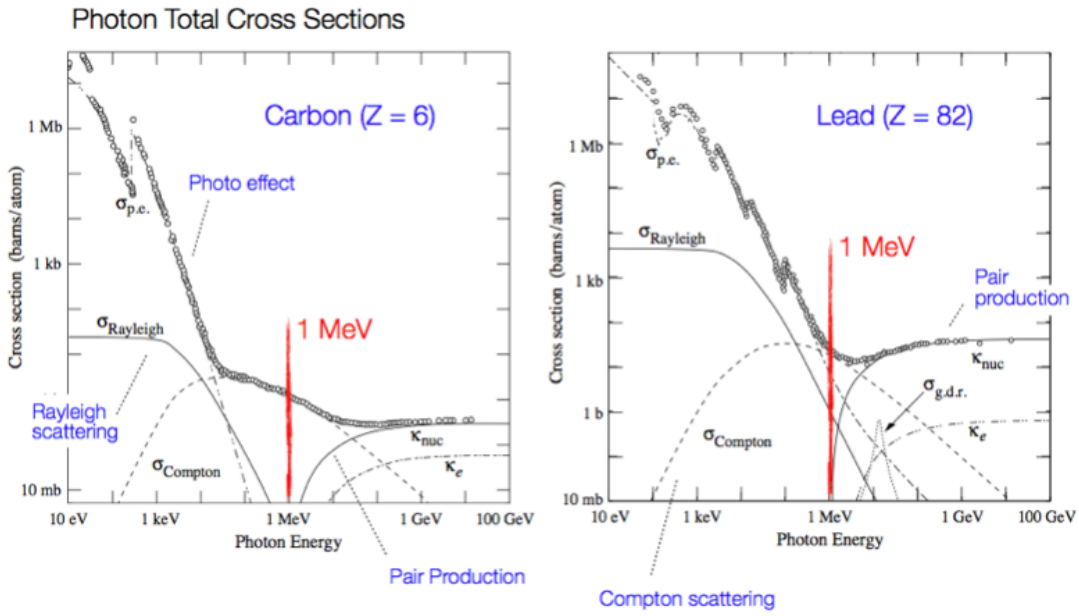


Fig. 7: Total cross-section of photons impinging on carbon (left) and on lead (right)

3.3 Cherenkov radiation

Charged particles passing through material at velocities larger than the speed of light in the material produce an EM shock-wave that materializes as EM radiation in the visible and ultraviolet range, the so-called Cherenkov radiation. With n being the refractive index of the material, the speed of light in the material is c/n. The fact that a particle does or does not emit Cherenkov radiation can then be used to apply a threshold to its velocity. The radiation is emitted at a characteristic angle with respect to the particle direction $\Theta_{\mathrm{c}} = \frac{c}{nv}$. Measuring the angle of the emitted light allows measurement of the particle velocity.

3.4 Transition radiation

Transition radiation is emitted when a charged particle crosses the boundary between two materials of different permittivities. The probability of emission is proportional to the Lorentz factor γ of the particle. It is only appreciable for ultra relativistic particles so it is mainly use to distinguish electrons from hadrons. As an example, a particle with $\gamma = 1000$ has a probability of about 1% to emit one photon at the transition between two materials but, by including many transition layers in the form of sheets, foam, or fibres, one can multiply the effect. The energy of the emitted photons is in the keV range.

4 EM and hadronic showers

4.1 Electromagnetic showers

One important consequence of the bremsstrahlung and pair-creation processes is the development of EM showers. EM calorimeters are therefore designed so that the shower development is contained and the deposited energy is collected.

For electrons, positrons, or photons of high energy (typically $E > 100$ MeV), showers result from cascading effects. Electrons undergo the bremsstrahlung process and emit photons and photons create a pair of electrons and positrons. This cascade continues until the emitted electrons are below the critical energy (E_{c} or ϵ_0). The number of ionization electrons or photons emitted by excited atoms is proportional to the energy of the incoming particle.

Figure 8(a) shows the number of electrons, as a function of the depth in units of X_0, in EM showers induced by electrons and photons of various energies (from Ref. [2]). Figure 8(b) shows the longitudinal energy loss profile for electrons and photons, comparing measurements and simulation (from Ref. [1]).

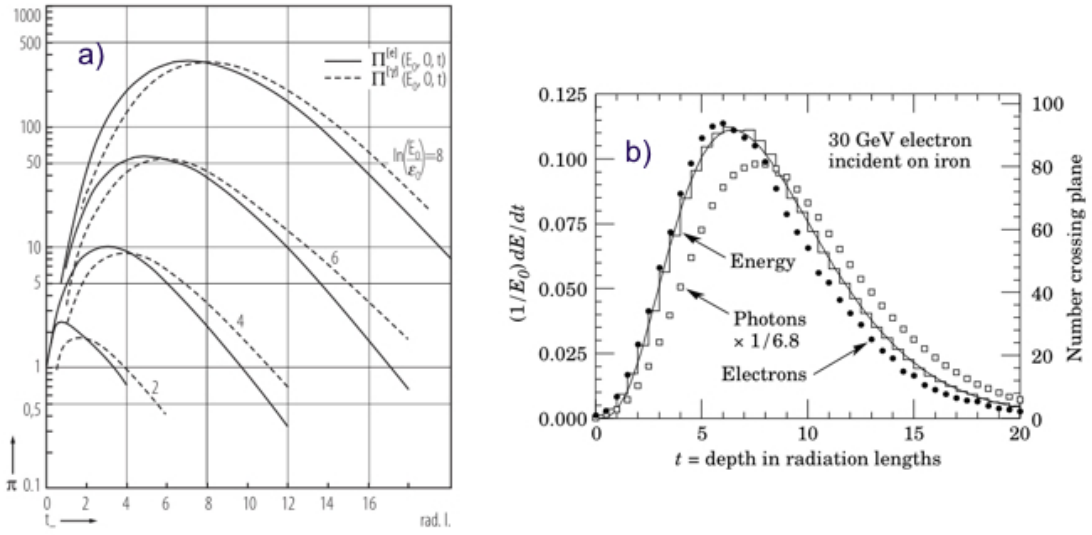


Fig. 8: (a) Number of electrons in electron and photon induced showers, for four energies, as a function of the depth in units of radiation length [9]. (b) Shower profile for 60 GeV incoming electrons in iron as a function of the depth t in unit of X_0 [1].

A description of the shower development has been proposed by Rossi [4] by computing the number of electrons plus positrons at a given energy and depth and similarly the number of photons. Accounting for the processes of bremsstrahlung, Compton scattering, ionization, and pair production, the authors have proposed a model which describes the shower development. The total track length (TTL) multiplied by the critical energy (E_c or ϵ_0) describes the energy transferred to the calorimeter medium by $\frac{\mathrm{d}E}{\mathrm{d}x}$, which constitutes the source of the calorimeter signal.

In such a model, the number of segment tracks increases as the depth, t, in units of radiation length as

$$N(t) = 2^t \tag{10}$$

and the average energy of each particle decreases as

$$E(t) = E_0/2^t \,, \tag{11}$$

until $E(t)$ reaches the critical energy. For $E(t) < E_\mathrm{c}$, ionization and excitation become dominant. In this model, the number of tracks is maximum at

$$t_\mathrm{max} = \ln \frac{E_0}{E_\mathrm{c}} / \ln 2 \,, \tag{12}$$

which has important implications in the context of detector design: that the shower depth varies as the logarithm of the particle energy. The shape and the energy dependence of the shower profile are presented in Fig. 9.

The TTL is therefore $E_0 \times X_0/E_\mathrm{c}$.

The higher the incident energy, the higher the TTL, and the better the relative resolution is. The shower development for electrons and photons differs by the shift of the start of the shower development of typically one radiation length.

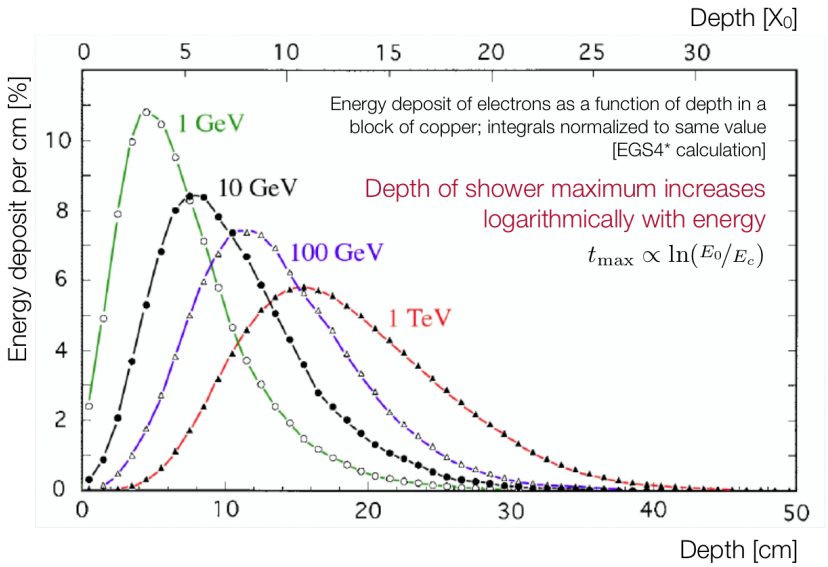


Fig. 9: Energy deposit of electrons with energies between 1 GeV and 1 TeV as a function of the depth in a block of copper.

4.2 Hadronic showers

Hadrons interact in matter dominantly via the nuclear interaction. By analogy with EM showers, the energy degradation of high-energy hadrons proceeds through an increasing number of (mostly) strong interactions with the calorimeter material. However, the complexity of the hadronic and nuclear processes produces a multitude of effects. The hadronic interaction produces two classes of secondary processes. First, energetic secondary hadrons are produced with momenta that are typically a fraction of the primary hadron momentum, i.e. at the GeV scale. Second, in hadronic collisions with the material nuclei, a significant part of the primary energy is diverted to nuclear processes such as excitation, nucleon evaporation, spallation, etc., resulting in particles with characteristic nuclear energies at the MeV scale. As an example, the energy deposition of a 5 GeV proton impinging on a block of lead and the scintillator can be decomposed as 40% ionization, 15% EM shower, 10% as carried by neutrons, 15% as photons from nuclear de-excitation, and 29% not detectable in the form of neutrinos or binding energy. This leads to less precision in energy resolution with respect to EM showers.

For high-energy incoming hadrons, the hadronic cross-section is fairly independent of energy and of the hadron type. The material dependence of the total hadronic inelastic cross-section on a material of mass A is given, in a simple form, by

$$\sigma_{\text{inelastic}}(p, A) \simeq \sigma_0 \times A^{0.7} \text{ with } \sigma_0 = 35 \text{ mb} , \tag{13}$$

where σ_0 is the inelastic cross-section of the proton–proton interaction.

One defines the interaction length by

$$\lambda_{\text{int}} = \frac{A}{N_{\text{A}} \sigma_{\text{inelastic}}(p, A)} \simeq 35 A^{1/3} \text{ g cm}^2 . \tag{14}$$

5 Energy loss transfer to detectable signals and signal collection

As presented in the preceding section, charged particles traversing matter create excited atoms, electron–ion pairs (in a gas or a liquid) or electron–hole pairs (in solid). This section summarizes existing techniques to exploit the photons emitted by excited atoms or the ionization.

5.1 Excitation

The photons emitted by the excited atoms can be detected with photon detectors such as photomultipliers or semi-conductor photon detectors. The emitted photons are typically in the range from ultraviolet to visible light. They are observed in noble gases (and even liquid), inorganic crystals, and organic scintillators. The principle is to convert the $\frac{\mathrm{d}E}{\mathrm{d}x}$ into visible light and detect the light with a photo-sensor. The typical light yield of scintillators is a few per cent of the energy loss. For instance, among the $\frac{\mathrm{d}E}{\mathrm{d}x} = 1.5$ MeV deposited in 1 cm of plastic scintillator of density $\rho = 1$ g cm^{-3}, 15 keV are available in the form of emitted photons and correspond to 15 000 photons.

The main features of photo-sensors are the sensitivity to energy, the fast time response, and the pulse shape discrimination. The requirements are high efficiency for conversion of the excitation energy to fluorescent radiation, transparency to the radiation to allow transmission of light, emission of light in a spectral range detectable for photo-sensors, and a short decay time to allow a fast response. The de-excitation time is an important parameter, in particular in the context of high-luminosity experiments such as at the Large Hadron Collider (LHC). For instance, in the $PbWO_4$ crystals of the CMS EM calorimeter, 80% of the light is emitted in 25 ns.

Two classes of scintillators are considered.

- **Inorganic crystals:** which are the substance with the largest light yield and are typically used for precision measurements of energetic photons but are typically slow.
- **Organic scintillators:** typically polycyclic hydrocarbons, such as naphthalene and anthracene, which typically have a lower light yield but a faster response than crystals. The light produced in the scintillator propagates to the edge where it is guided in light-guides, with total reflection, to the detector device. In addition, the use of a wavelength shifter, converting the light to a higher wave length, allows the photon to be transported without reflecting back to the scintillator.

The classical device used to convert these photons into electrical signals is the photo-multiplier. A photon hits a photo-cathode, a material with a very small work function, and liberates an electron which is then accelerated in a strong electric field to a dynode, made of a material with high secondary electron yield. The one original electron will therefore create several electrons, which are again guided to the next dynode, and so on. Out of one single electron one ends up with a sizable signal typically of 10^7–10^8 electrons. In recent years, the use of solid-state photo-multipliers, such as avalanche photo-diodes, vacuum photo-diodes, and silicon photo-multipliers, has become popular as they are insensitive to magnetic fields and less expensive.

5.2 Ionization

By applying an electric field in the detector volume, the ionization electrons and ions are moving, which induces a signal on metallic electrodes. These signals are then readout by appropriate readout electronics.

The noise and pre-amplifier determine whether the signal can be registered. The signal-to-noise ratio must be large: $S/N \gg 1$. The noise is characterized by the Equivalent Noise Charge (ENC) which is the charge at the input that produces an output signal equal to the noise. The ENC of very good amplifiers can be as low as 50 e^-, a typical value being 1000 e^-. In order to register a signal, the registered charge must be $q \gg$ ENC, i.e., typically $q \gg 1000\ e^-$. For a gas detector, $q \simeq 80\ e^-/$ cm is too small to be detected. Solid-state detectors have 1000 times more density and a factor of 5–10 less ionization energy. Therefore, the primary charge, in a solid-state detector, therefore reaches 10^4–10^5, which is the same for a gas detector.

5.2.1 Gas detectors

Gas detectors need internal amplification in order to be sensitive to a single particle. The amplification processes and drift in an electric field are the basis of the operation of gas chambers. Ionization detectors

are generally operated in the proportional regime where an amplification of 10^4 to 10^6 is used. The amplification of the signal in gas is schematically represented in Fig. 10. In the case of a cylindrical geometry, the amplification process can be described by the electric field $E(r)$ as a function of r, the radial distance between the charge particle and the anode wire, and the potential $V(r)$,

$$E(r) \propto \frac{1}{r} \quad \text{and} \quad V(r) \propto \ln \frac{r}{a} \, .$$

The primary electrons drift towards the positive anode. Close to the very thin wire, due to the $1/r$ dependence, the electric field reaches values $E > 1$ kV/ cm
(top-left curve of Fig. 10). In between collisions with atoms, electrons gain enough energy to ionize further gas molecules generating an exponential increase in the number of electron–ion pairs close (a few μm) to the wire (bottom of Fig. 10).

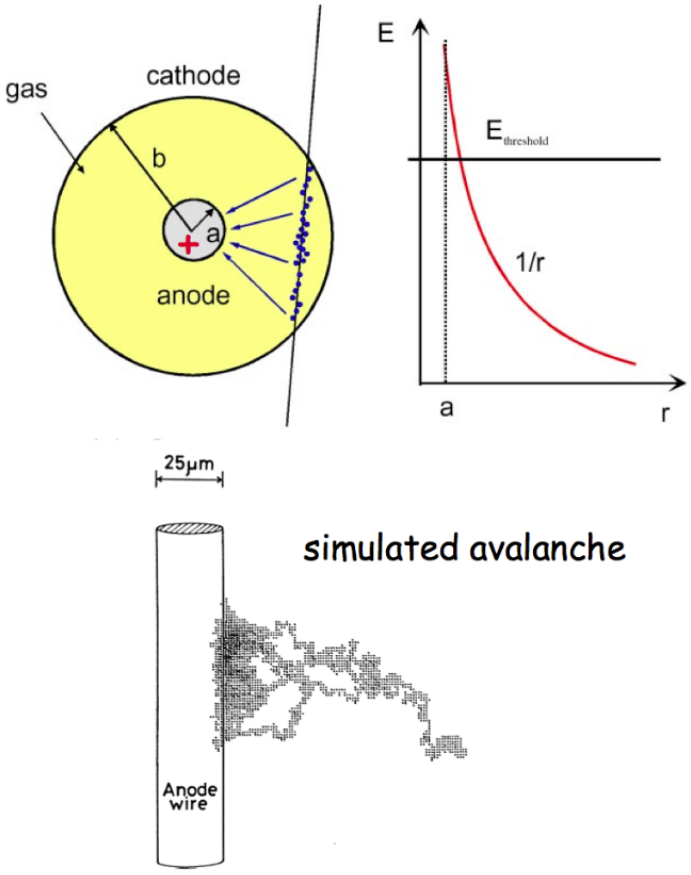


Fig. 10: Top-left: Schematic representation of a gas drift tube. Top-right: Electric field dependence with r. Bottom: Simulation of an ionization avalanche onto an anode wire of diameter 25 μm.

Gas detectors are most often used as tracking detectors in order to reconstruct the charged particle trajectory and measure its momentum from its curvature induced by the magnetic field. Criteria to obtain an optimal momentum resolution are to have many measurement points, a large detector volume, very good single point resolution, and as little multiple scattering as possible.

The response of a proportional chamber is shown in Fig. 11 as a function of the applied voltage. There are several distinctive regions of the response curve: the ionization regime where the primary charge is collected, giving a flat response; the proportional regime where the electric field is large enough to generate multiplication, with a gain up to 10^6; and the Geiger–Muller regime, where strong photon emission propagates avalanches.

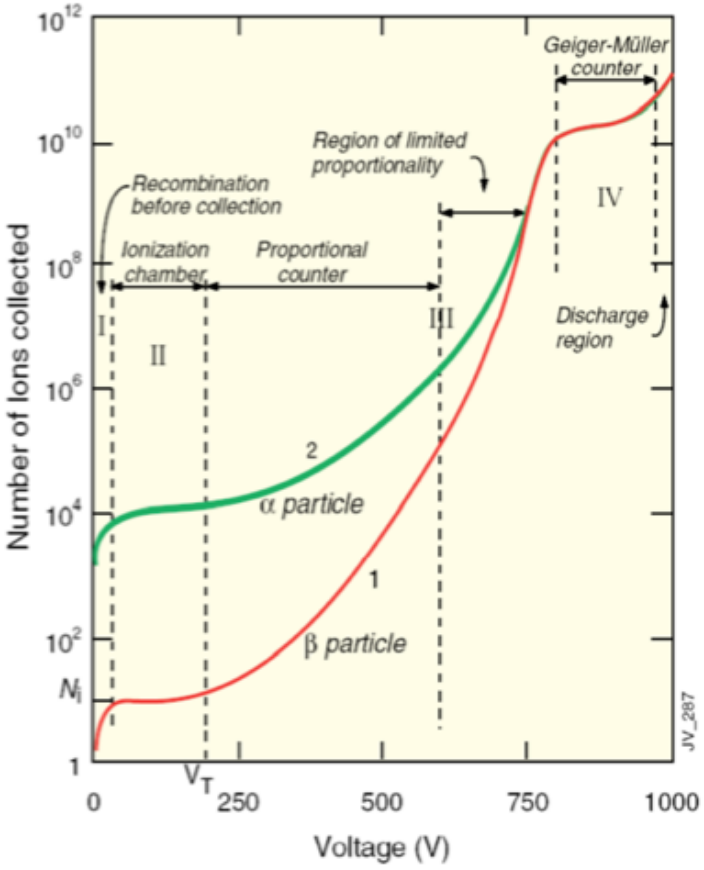


Fig. 11: Number of ions collected as a function of the applied voltage and definition of the operation regimes

The proportional mode is most often exploited with wire chambers. Wires of very small diameter, between 10 μm and 100 μm, are placed between two metallic plates a few millimeters apart. The wires are at a high voltage (a few kV), 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 field close to the wires, the electrons are accelerated to energies above the gas ionization energy, which results in the production of secondary electrons and as a consequence an electron avalanche.

The position of the primary ionization electrons can be determined by segmenting the cathodes (metal plates) into strips, which are sensitive to the induced charge. Another way to achieve position resolution, far smaller that the separation between the wires, is to measure the drift time of the charge with respect to a reference clock such as an accelerator clock. For example, the precision on the track position achieved by the ATLAS muon system is of order 80 μm for a distance between the wires of order 15 mm.

Gas chambers have been extensively used in HEP detectors. They provide many measurement points, large volume, very good single point resolution, and as little multiple scattering as possible. They are perfectly suited for the large detector areas at outer radii. They are not suited to the small radius of LHC experiments for instance, where the particle flow is large. Several dedicated techniques have been employed such as multiwire proportional chambers, drift chambers, time projection chambers, streamer tubes, and resistive plates chambers.

In the last 10–15 years, a large variety of new gas detectors have been developed in the context of particle physics instrumentation, such as micro-pattern gas detector á la GEM (a gas electron multiplier) or the Micromegas (micro-mesh gas detector).

5.2.2 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 parent atom. 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 in 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 a semi-conductor.

In order to use a semi-conductor 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 whereas silicon and germanium must be cooled or the free charge carriers must be eliminated by other tricks such as doping.

The diamond detector works like a solid-state ionization chamber. A diamond of a few hundred micrometres thickness is placed between two metal electrodes and an electric field is applied. The very large electron and hole mobilities of diamond result in very fast and short signals. Diamond is therefore both a tracking and a timing detector. Small diamond detectors are installed as a beam condition monitor, a few centimetres from the beam pipe, in the ATLAS detector.

Silicon is the most widely used semiconductor material for particle detection. A high-energy particle produces 33 000 electron–hole pairs per cm^3. However, at room temperature there are 1.45×10^{10} electron–hole pairs per cm^3. To be able to operate a silicon detector at room temperature, doping is employed. Doping silicon with arsenic makes it an n-type conductor (more electrons than holes) whereas doping silicon with boron makes it p-type conductor (more holes than electrons). Putting an n-type and p-type conductor in contact creates a diode.

At a p–n junction, the charges are depleted and a zone that is 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 induce an electrical signal on the metal electrodes. As silicon is the most commonly used material in the electronics industry, this constitutes a big advantage with respect to other materials.

Strip detectors are a very common application, where the detector is segmented into strips of a few 50–150 μm pitch and the signals are read out on the ends by wire bonding the strips to the readout electronics. The other co-ordinate can then be determined, either by another strip detector with perpendicular orientation or by implementing perpendicular strips on the same wafer.

In the very high multiplicity region close to the collision point, a pixel detector for sizes a few tens of micrometres by a few hundreds of micrometres can be used. The readout of a pixel module is achieved by building the readout electronics wafer in the same geometry as the pixel layout and soldering, via bump bonding, each of the pixels to its respective amplifier. Pixel systems of about 100 million channels are successfully operating at LHC. The typical vertex resolution achieved is approximately 30 μm.

Current developments in the solid-state detector domain are exploration of the possibility to integrate the detector element and the readout electronics, as well as the application of CMOS (complementary metal-oxide semiconductor) sensors.

6 Calorimeters

Calorimeters measure the energy of neutral and charged particles by complete absorption.

The principle is to measure the signal, induced by electrons and positrons with energy below the critical energy, which is proportional to the incident energy. Calorimeters vary by the technique used to collect the signal (sensitive material) and by the technique to induce the shower development (passive material). Two times two general types of calorimeters have been built: EM (Section 6.3) or hadronic calorimeters (Section 6.4) on one dimension and homogeneous (Section 6.1) or sampling calorimeters (Section 6.2) on the other dimension.

Calorimeters are a natural complement to tracking detectors as they measure the energy of both neutral and charged particles. In calorimeters, the relative energy resolution improves with energy because it is governed by a statistical process. In contrast, relative momentum resolution degrades with energy for tracking detectors in a magnetic field.

6.1 Homogeneous calorimeters

A homogeneous calorimeter is built only from the sensitive medium. In principle, for a similar containment and signal detection efficiency, a homogeneous calorimeter gives the best energy resolution because sampling calorimeters are limited by *sampling fluctuation*.

The calorimeter energy resolution is determined by fluctuations, such as shower fluctuations, photo-electron statistics, and shower leakage, and instrumental effects such as noise. Accounting for these limitations, the relative energy resolution of a calorimeter can be written as

$$\frac{\sigma(E)}{E} = \frac{a}{\sqrt{E}} \oplus \frac{b}{E} \oplus c\% \,, \tag{15}$$

where a is called the statistical or sampling term, b the noise term, and c the constant term.

For an ideal (homogeneous) calorimeter without leakage, the energy resolution is limited only by statistical fluctuations of the number N of shower particles, i.e.

$$\frac{\sigma(E)}{E} \propto \frac{\sigma(N)}{N} \approx \frac{\sqrt{N}}{N} = \frac{1}{\sqrt{N}} \quad \text{with} \quad N = \frac{E}{W} \,, \tag{16}$$

with E being the energy of the incoming particle and W the mean energy required to produce a *signal quantum*. For instance $W \approx$3.6 eV for a silicon detector, 30 eV for gas detectors and 100 eV for a plastic scintillator. The formulation of the relative energy resolution of Eq. (16) needs to be corrected to account for correlations between fluctuations; the correction factor, F, is called the Fano factor [2]:

$$\frac{\sigma(E)}{E} \propto \sqrt{\frac{FW}{E}} \,. \tag{17}$$

Homogeneous calorimeters are based on three main primary signal collection: scintillation light ($PbWO_4$, BGO, BaF_2), Cherenkov light (lead glass), and ionization signal (in noble gases such as argon, krypton, and xenon).

The CMS EM calorimeter is built of ≈70 000 lead tungsten crystal. The energy resolution, for the barrel part of the CMS EM calorimeter [6], reaches 1.1% for a non-converted photon with $E_\perp^\gamma \simeq 60$ GeV, and is about 1.5% for electrons with $E_\perp^{\mathrm{e}} \simeq 45$ GeV from the Z^0 decay with low bremsstrahlung. Because of inter-calibration, the measured constant term varies between 0.3% and 0.5% for the barrel part of the calorimeter and between 1% and 1.5% in the end-cap, depending on the pseudo-rapidity.

The NA48 EM calorimeter is a homogeneous liquid krypton calorimeter. The energy resolution for photons has been measured to be

$$\frac{\sigma_{\mathrm{e}}}{E} = \frac{(3.2 \pm 0.2)\%}{\sqrt{(E)}} \oplus \frac{0.09 \pm 0.01}{E} \oplus (0.42 \pm 0.05)\% \,,$$

and the energy linearity is better than $\approx 0.1\%$ for electrons in the energy range 5–100 GeV. The linearity was measured using the $\frac{E}{p}$ technique, where the energy, E, measured by the calorimeter is compared to the particle momentum measured from the spectrometer.

6.2 Sampling calorimeters

A sampling calorimeter consists of plates of dense, passive material alternating with layers of sensitive material. For EM showers, passive materials with low critical energy (thus high Z) are used, thus maximizing the number of electrons and positrons in a shower to be sampled by the active layers. In practice, lead is most frequently used. The thickness, t, of the passive layers (in units of X_0) determines the sampling frequency, i.e. the number of times a high-energy electron or photon shower is *sampled*. Intuitively, the thinner the passive layer (i.e. the higher the sampling frequency), the better the resolution should be. The thickness, u, of the active layer (in units of X_0) is usually characterized by the sampling fraction f_S defined as

$$f_S = \frac{u \times \frac{\mathrm{d}E}{\mathrm{d}x}(\text{active})}{u \times \frac{\mathrm{d}E}{\mathrm{d}x}(\text{active}) + t \times \frac{\mathrm{d}E}{\mathrm{d}x}(\text{passive})} \tag{18}$$

where u, t are in g cm^{-2} and $\frac{\mathrm{d}E}{\mathrm{d}x}$ is in MeV g^{-1} cm^2.

This *sampling* of the energy results in a loss of information and hence in additional sampling fluctuations. An approximation [2] for these fluctuations in EM calorimeters can be derived using the TTL (see Section 4.1) of a shower, initiated by an electron or photon of energy E.

The signal is approximated by the number N_x of e^+ or e^- traversing the active signal planes, spaced by a distance $(t + u)$.

This number N_x of crossings is

$$N_x = \frac{TTL}{t+u} = \frac{E}{E_\mathrm{c}} \frac{1}{t+u} \,. \tag{19}$$

Assuming statistical independence of the crossings, the fluctuations in N_x represent the *sampling fluctuations* $\sigma(E)_{\text{samp}}$,

$$\frac{\sigma(E)_{\text{samp}}}{E} = \frac{\sigma(N_x)}{N_x} = \frac{1}{\sqrt{N_x}} = \frac{a}{\sqrt{E[\,\text{GeV}]}} \,. \tag{20}$$

As an example, Fig. 12 shows a schematic view of the barrel section of the ATLAS EM calorimeter which is built with lead plates (e.g. 1.53 cm thick for the central part), interleaved with copper electrodes to build a 6.4 m cylinder of 2.8 m inner diameter and 4 m outer diameter. The lead absorbers and the electrodes are folded with an accordion shape which allows signal collection on the front and back faces of the calorimeter. This has two main advantages: a fast signal collection and the absence of gaps between cells, leading to a complete coverage in the azimuthal direction. Signal pads are drawn on the electrodes to measure the particle position and sample the shower development. The cylinder is housed in a cryostat filled with liquid argon (LAr), and the lead constitutes the passive material and LAr the active material. The ionization electrons drift towards the electrodes through which a high voltage is applied. A current is induced which is driven to preamplifiers located outside of the cryostat.

The energy resolution, for the barrel part of the ATLAS EM calorimeter [7], is 1.5% for non-converted photons with $E_\perp^\gamma \simeq 60$ GeV, and is about 1.5% for electrons with $E_\perp^\mathrm{e} \simeq 45$ GeV from the Z^0 decay. The measured constant term varies between 0.7% and 1% for the barrel part of the calorimeter and between 1% and 2.5% in the end-cap, depending on the pseudo-rapidity.

6.3 EM calorimeters

EM showers develop as described in Section 4.1. The number of particles in the shower increases until the average energy of the produced particle is below the critical energy when no more particles can

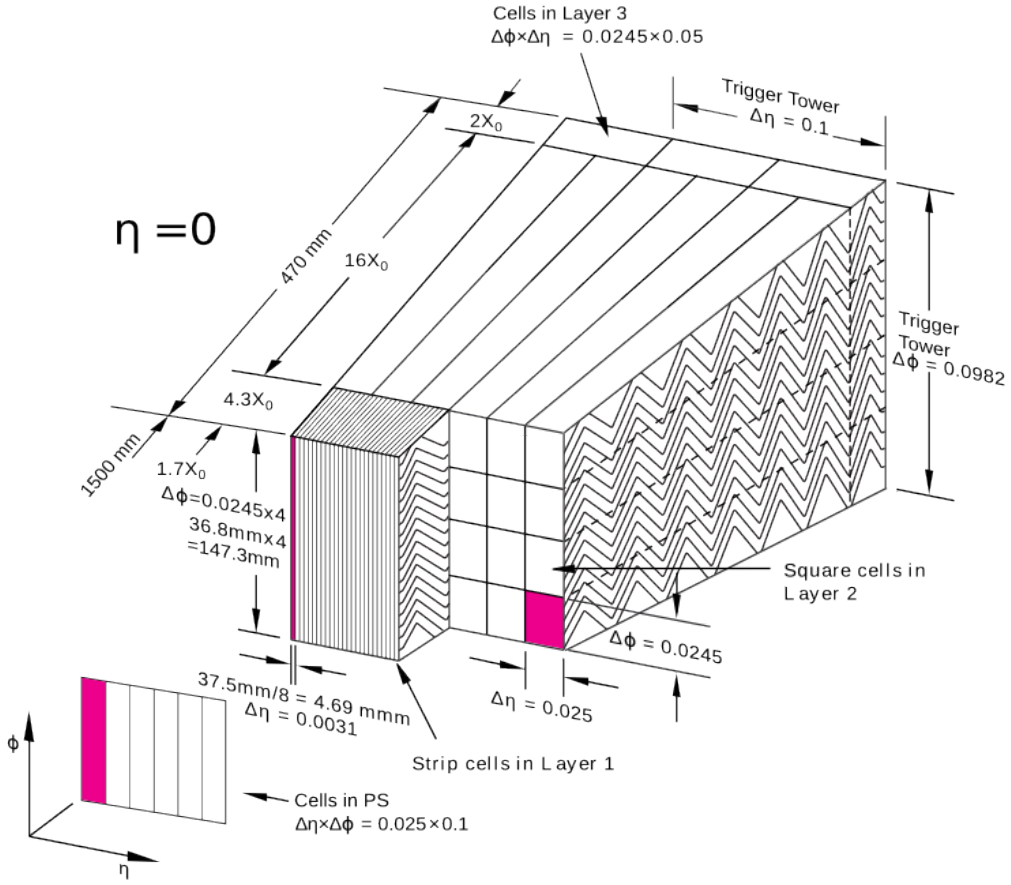


Fig. 12: Schematic representation of the ATLAS liquid argon calorimeter

be produced. The particles in the shower will ionize the medium or undergo Compton scattering. The lateral development of the shower is mainly governed by the electrons that do not radiate but have enough energy to travel far away from the axis.

The typical relative energy resolution for sampling EM calorimeters at the LHC is

$$\frac{\sigma(E)}{E} = \frac{10\%}{\sqrt{E[\,\text{GeV}]}} \oplus \frac{100\ \text{MeV}}{E[\,\text{GeV}]} \oplus 0.5\text{–}1\% \,. \tag{21}$$

6.4 Hadronic calorimeters

The development of hadronic showers is more complex than the development of EM particles (see Section 4.2). The fraction of detectable energy from hadronic showers is smaller than for EM showers, leading to an intrinsically worse relative energy resolution for hadrons than for electrons and photons. In order to contain high-energy hadrons, hadronic calorimeters need to be larger than EM calorimeters. As an example, the interaction length λ_{int}, which describes the typical size of one nuclear interaction, is 17 cm in iron when the radiation length X_0 is 1.7 cm. In addition, as hadronic showers extend deeper and wider than EM showers, the granularity of a hadronic calorimeter is coarser than for an EM calorimeter. Hadron calorimeters are mainly sampling calorimeters, with iron or copper as the passive material and a scintillator as the active material. The typical relative energy resolution of hadronic calorimeters for the LHC is

$$\frac{\sigma(E)}{E} = \frac{50\text{–}100\%}{\sqrt{E[\,\text{GeV}]}} \,. \tag{22}$$

7 Particle identification

Particle identification is typically the result of the combination of several observations from various instruments. Only a few examples are given here.

By combining the energy loss along a charged track, which is a function of the velocity β, with the momentum measurement from the curvature in the magnetic field, one can extract the particle mass, providing the momentum is small enough. This is illustrated by Fig. 13 from Ref. [5].

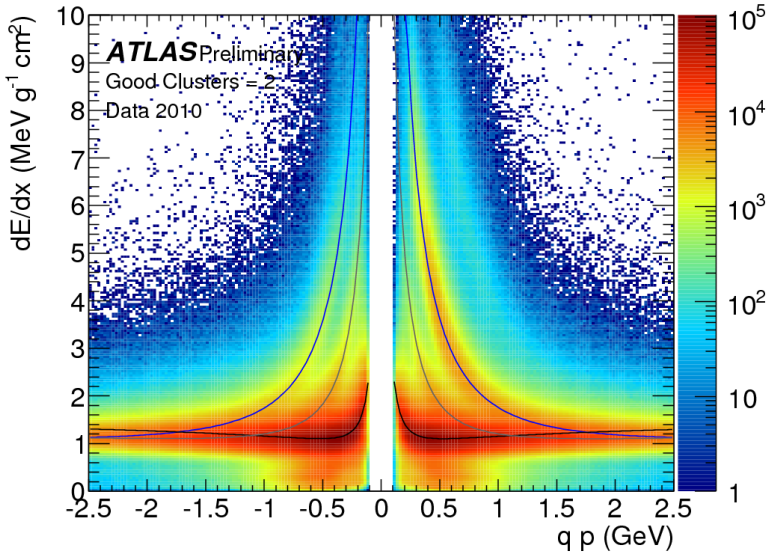


Fig. 13: Bi-dimensional distribution of $\frac{\mathrm{d}E}{\mathrm{d}x}$ and momentum for ATLAS 2010 data. The distributions of the most probable value for the fitted probability density functions of pions (black), kaons (grey), and protons (blue), in different track categories, are superimposed.

The ATLAS EM calorimeter segmentation allows identification of single photons from π^0 decays to two photons by its ability to separate single from double showers in the thinly segmented first layer. The is illustrated in Fig. 14. The resulting rejection power of the calorimeter for π^0 is about three for high-energy isolated photons.

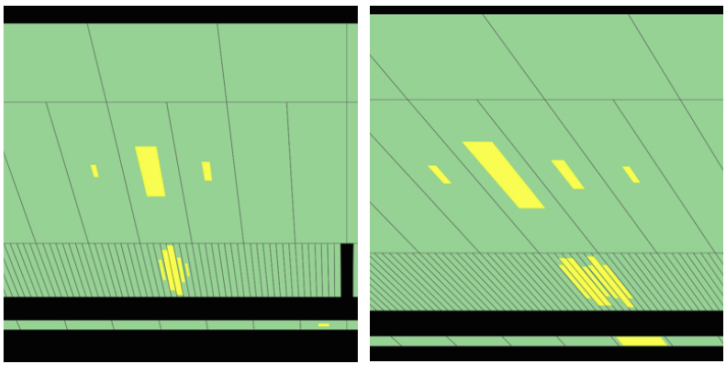


Fig. 14: Energy depositions in the successive layers of the ATLAS EM calorimeter of a candidate photon on the left and a candidate π^0 on the right.

As illustrated by Fig. 1, the association of the tracks measured in the inner tracking system with the energy deposits in the calorimeter makes it possible to distinguish between electrons and photons. The shape of the energy deposition in the EM and hadronic calorimeters makes it possible to distinguish

between electrons and hadrons.

8 Detectors

Additional components are necessary to build and operate a complete detector. In the previous sections, only a few topics have been discussed. Many more would need to be included to give a fair description of the versatility, complexity and refinement of the design and operation of a detector. Among the missing topics are: low noise electronics, fast digital electronics, selective and efficient trigger systems, high data flow systems, real-time software, highly radiation tolerant components, reconstruction software, and calibration.

Revealing and measuring rare phenomena are the main goals for LHC physics. This imposes extremely severe constraints on the four main LHC detectors: ATLAS, CMS, LHCb, and ALICE. These four detectors are large, complex, and have very high capability. Huge magnet systems dominate their mechanical structures. The proton collision rate of 1 GHz produces particles and jets of TeV-scale energy. This imposes severe demands in terms of spectrometer and calorimeter size, rate capability, and radiation resistance. The fact that a few hundred events, out of the 10^9 events produced every second, can be written to disk necessitates highly complex online event selection as a trigger.

The basic layout of LHC collider experiments is quite similar.

- **Tracking detector:** charged particles are bent inside a solenoidal magnetic field produced by the magnet surrounding the tracking system. The charged particle momentum is reconstructed from the association of hits collected in the sensitive tracking detectors along the track path.
  - **Vertex detector:** close to the interaction point, there are several layers of pixel detectors which allow the position of charged particle vertices to be measured to a few tens of micrometres. This also allows short-lived B and D mesons to be identified.
  - **Spectrometer:** to follow the track curvature a succession of sensitive layers (mainly silicon strips) are positioned around the interaction region and typically up to 1 m in radius and a few metres along the beam line. The CMS tracking detector is built from silicon sensors only, whereas the other experiments use silicon at low radius and a gas detector further along the track path.
- **Calorimeters:** the tracking detector is surrounded by the EM and hadronic calorimeters, which measure the energy of electrons, photons, hadrons, and jets by absorbing them. The hermiticity of the calorimeters in the transverse direction and down to very small angles close to the beam line allows reconstruction of the total transverse energy deposited with a high-performance resolution thereby allowing reconstruction of the transverse energy carried away by neutrinos or any non-interacting particle. The LHC calorimeters are highly segmented in order reconstruct the position of neutral particles, to separate electrons from hadrons, and to be less sensitive to the energy pile-up from simultaneous proton–proton interactions.
- **Muons systems:** muons, which deposit very little energy in the calorimeters and are therefore not stopped, are measured at very large radii by dedicated muons systems.

The sequence of the vertex detector, spectrometer, calorimetry, and muon detector is the classic basic geometry that underlies most collider and fixed-target experiments. There are many other types of detectors, from very small to very large arrays of telescopes, for example. Each detector has its own specialism such as neutrino detection or detection of very high-energy gamma rays from astrophysical sources. These are only a few examples from a large variety of existing detector systems. However, it is important to remember that there are only a few basic principles of particle interaction with matter that underlie these different detectors.

Acknowledgements

I have profited enormously, in the preparation of the lectures and of these proceedings, from the material prepared by Hans Christian Schultz-Coulon [12], Marco Delmastro [10], Daniel Fournier [9], and Werner Riegler [11].

References

[1] K.A. Olive *et al*. (Particle Data Group), *Chin. Phys.* C **38**(9) (2014) 090001, http://dx.doi.org/10.1088/1674-1137/38/9/090001

[2] C.W. Fabjan and D. Fournier, *Detectors for Particles and Radiation. Part 1: Principles and Methods*, Eds. C.W. Fabjan and H. Schopper (Springer, Berlin, 2011), Vol. 21B1 of the series Landolt-Bernstein – Group I Elementary Particles, Nuclei and Atoms, pp. 145–193, http://dx.doi.org/10.1007/978-3-642-03606-4

[3] C.W. Fabjan, in *Experimental Techniques in High Energy Physics*, Ed. T. Ferbel (Addison Wesley, Menlo Park, CA, 1987).

[4] B. Rossi and K. Greisen, *Rev. Mod. Phys.* **13**(4) (1941) 240, http://dx.doi.org/10.1103/RevModPhys.13.240

[5] The ATLAS Collaboration, dE/dx measurement in the ATLAS Pixel Detector and its use for particle identification, ATLAS-CONF-2011-016 (2011).

[6] The CMS Collaboration, Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV, *J. Instrum.* **8** (2013) P09009. http://dx.doi.org/10.1088/1748-0221/8/09/P09009

[7] The ATLAS Collaboration, Electron and photon energy calibration with the ATLAS detector using LHC Run1 data, *Eur. Phys. J.* C **74** (2014) 3071, http://dx.doi.org/10.1140/epjc/s10052-014-3071-4

[8] G. Unal for the NA48 Collaboration, Performances of the NA48 Liquid Krypton Calorimeter, Proceedings CALOR2000 Conference, Annecy, 2000, https://cds.cern.ch/record/479427

[9] D. Fournier, Lecture at EDIT 2011 School on Instrumentation, https://indico.cern.ch/event/124299/

[10] M. Delmastro, Lecture at European School of Instrumentation for Particle and Astroparticle Physics (ESIPAP), https://indico.cern.ch/event/294651/

[11] W. Riegler, Particle Physics Instrumentation, Proceedings of the First Asia-Europe-Pacific School of High-Energy Physics, Fukuoka, Japan, 14 - 27 Oct 2012, http://dx.doi.org/10.5170/CERN-2014-001.241

[12] H.C. Schultz-Coulon, The Physics of Particle Detectors, Lecture Notes, University of Heidelberg, http://www.kip.uni-heidelberg.de/~coulon/Lectures/Detectors/

Local Organizing Committee

Subhasis Chattopadhyay (VECC)
Rohini Godbole (IISc, Chair)
Gobinda Majumdar (TIFR)
Prolay K. Mal (NISER)
Sreerup Raychaudhuri (TIFR)
Pradip K. Sahu (IOP)

Advisors to the Local Organizing Committee

Mustansir Barma (TIFR)
Ajit K. Mohanty (BARC)
Naba K. Mondal (TIFR)
Tapan K. Nayak (VECC)
Sudhakar Panda (IOP)
Sibaji Raha (Bose Institute)

International Organizing Committee

Mark Boland (Australian Synchrotron)
Subhasis Chattopadhyay (VECC)
Simon Eidelman (BINP)
Nick Ellis (chair) (CERN)
Kazunori Hanagaki (Osaka University)
Yee Bob Hsiung (National Taiwan University)
Pyungwon Ko (KIAS)
Martijn Mulders (CERN)
Sreerup Raychaudhuri (TIFR)
Lydia Roos (IN2P3/LPNHE Paris)
Thomas Schoerner-Sadenius (DESY) - from March 2014
Didier Vilanova (CEA/Irfu)
Changzheng Yuan (IHEP)
Shi-Lin Zhu (Peking University)

International Advisory Committee

Etienne Augé (IN2P3/CNRS) - until September 2014
Ursula Bassler (IN2P3/CNRS) - from September 2014
Alexander Bondar (BINP) - from March 2014
Mikhail Danilov (ITEP) - until March 2014
Rohini Godbole (IISc)
Shih-Chang Lee (Academia Sinica) - until September 2014
Joachim Mnick (DESY)
Mitsuaki Noazki (EK, Chair) - from March 2014
Martin Sevior (University of Melbourne)
Xiaoyan Shen (IHEP) - from March 2014
Dongchul Son (KNU)
Fumihiko Takasaki (chair) (KEK) - until March 2014
Henry Tsz-king Wong (Academia Sinica) - from September 2014
Rüdiger Voss (CERN)
Yifang Wang (IHEP) - until March 2014

Lecturers

Sourendu Gupta (TIFR)
Koichi Hamaguchi (U. Tokyo)
Rolf Heuer (CERN)
Seung J. Lee (KAIST)
Valery Rubakov (INR Moscow)
Peter Skands (CERN and Monash U.)
Atsuto Suzuki (KEK)
Wouter Verkerke (Nikhef)
Isabelle Wingerter-Seez (IN2P3/LAPP)
Zhi-Zhong Xing (IHEP)

Discussion Leaders

Aleksandr Azatov (CERN)
Monika Blanke (CERN and KIT)
Cheng-Wei Chiang (National Central U.)
Ian-Woo Kim (CERN)
Sudhir Vempati (IISc)

Students

Subash ADHIKARI

Sk Noor ALAM

Md Hafizuddin AL-HELMY NOUXMAN

Debjyoti BARDHAN

Neil BARRIE

Alexey BASKAKOV

Inayat BHAT

Sandeep BHOWMIK

Alexander BYLINKIN

Indrani CHAKRABORTY

Li-Chu CHANG

Kalyanmoy CHATTERJEE

Subikash CHOUDHURY

Dipankar DAS

Ram Krishna DEWANJEE

Sourav DEY

Tatiana DROZHZHOVA

Nadine FISCHER

Rajesh GANAI

Reza GOLDOUZIAN

Deepanjali GOSWAMI

Lucia GRILLO

Chandan GUPTA

Ruchi GUPTA

Kouhei HANZAWA

Yusho HOMMA

Aliaksei HRYNEVICH

Kuo-Lun JEN

Raveendrababu KARANAM

Rose KOOPMAN

Ievgen KOROL

Anastasiia KOZACHUK

Prasanth KRISHNAN KP

Fabian KUGER

Dinesh KUMAR

Andrey KUPICH

Jayita LAHIRI

Robyn LUCAS

Bibhuprasad MAHAKUD

Timofey MALTSEV

Soureek MITRA

Md. MOHSIN

Simon Stark MORTENSEN

Yuki NAKAI

Ryutaro NISHIMURA

Genessis PEREZ RIVERA

Sylvestre PIRES

Soumita PRAMANICK

Narayan RANA

Thomas RAVENSCROFT

Norma RISDIANTO

Ashim ROY

Ipsita SAHA

Dibyakrupa SAHOO

Niladribihari SAHOO

Artur SHAIKHIEV

Ralitsa SHARANKOVA

Avirup SHAW

Weimin SONG

Marian STAHL

Nairit SUR

Tatsuhiko TOMITA

Nam TRAN

Sijing ZHANG

Posters

Poster title	Presenter
Search for anomalous single top quark production in association with a photon in pp collisions at 8 TeV	REZA GOLDOUZIAN
Development of ultra cold muon source for muon g-2/EDM experiment at J-PARC	KOUHEI H
Search for anomalous Wtb couplings and top FCNC in t-channel single top quark events	BASKAKOV, A
Study of hadronic event-shape variables in multijet final states in pp collisions at 7 TeV	BHOWMIK, S; MAJUMDER, G; BANERJEE, S; ROY, D; MAITY, M; GUCHAIT, M
Bottom-up with Higgs Naturalness	INDRANI CHAKRABORTY
Search for the Standard Model Higgs boson in pp collisions using the CMS detector in association with a W boson where W $\to \ell\nu$ and H $\to \tau$ H+τ H	CHATTERJEE K
Searches for SUSY with Compressed Mass Spectra at CMS	LUCAS, R
The Higgs or 'a Higgs'?	DAS, DIPANKAR
Study of Additional Scalars and Fermions in Models beyond the Standard Model	GOSWAMI, D
Measurement of the semileptonic CP asymmetry in $B^0 - \overline{B^0}$ mixing	GRILLO, L.
Statistical approach in quantifying sensitivity of the experiments	GUPTACHANDAN
Calibration of CMS Hadron Calorimeter	GUPTA, R.
Measurement of three-jet production cross-section in pp collisions at 7 TeV using the ATLAS detector	HRYNEVICH, A
Neutrino Mass and Dark Matter in the light of AMS-02 Result	SAHA, IPSITA

Poster title	**Presenter**
Fragmentation fractions	KOOPMAN, R.F.
New Physics in $e^+e \to \gamma Z$ at the ILC with polarized beams: explorations beyond conventional anomalous triple gauge boson couplings	LAHIRI JAYITA
Investigation and optimization of the detectors based on gas electron multipliers	MALTSEV, T.V.
Prospects of measuring asls in LHCb	STAHL, M.
Development for Silicon-On-Insulator Pixel detectors	NISHIMURA, R.
Measurement of electromagnetic form factor of the pion with SND on VEPP-2000	KUPICH A.
Inclusive Jet Cross Section Measurement in LHC at 14 TeV	DEY, S.
Study of Pileup Removal Algorithm for Jets MAHAKUD	BIBHUPRASAD
Inflationary Baryogenesis with Gauged Baryon Number	BARRIE, N.
Development of Large Strip RPCs as TOF system	TRAN NAM, TOMIDA NATSUKI
Search for Higgs Bosons in Standard Model and MSSM at CMS Detector	RAM KRISHNA DEWANJEE
Study of float glass as electrode for RPC	RAVEENDRABABU, K; BEHERA, P. K; SATYANARAYANA, B; MUKHOPADHYAY, S; MAJUMDAR, N
Neutrino Mass and Dark Matter in the light of AMS-02 Result	SAHA, IPSITA
Exploring violations of Bose, CP and CPT symmetries via Dalitz plots and Dalitz prisms	SAHOO, D.
Neutrino Oscillation Parameters: Are they all related?	PRAMANICK, S; RAYCHAUDHURI,A
Measurement of the top quark mass in leptonic decays of t-channel single top events at 8 TeV	MITRA, S