CERN Yellow Reports: Monographs CERN-2025-009

Proceedings of the 2023 European School of High-Energy Physics

Grenaa, Denmark, 6–19 September 2023

Editors: Sascha Stahl, Alexander Huss

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

ISBN 978-92-9083-714-5 (paperback)
ISBN 978-92-9083-715-2 (PDF)
ISSN 2519-8041 (Print)
ISSN 2519-805X (Online)
DOI https://doi.org/10.23730/CYRSP-2025-009

This volume should be cited as:

Proceedings of the 2023 European School of High-Energy Physics,
S. Stahl, A. Huss (eds.)
CERN Yellow Reports: School Proceedings, CERN-2025-009 (CERN, Geneva, 2021)
https://doi.org/10.23730/CYRSP-2025-009.

A contribution in this report should be cited as:

[Chapter editor name(s)], in Proceedings of the 2023 European School of High-Energy Physics, S. Stahl and A. Huss (eds.)
CERN-2025-009 (CERN, Geneva, 2025), pp. [first page]–[last page],
https://doi.org/10.23730/CYRSP-2025-009.[first page]

Corresponding editor: Sascha.Stahl@cern.ch.

Accepted in Dec. 2025, by the CERN Reports Editorial Board (contact Carlos.Lourenco@cern.ch).

Published by the CERN Scientific Information Service (contact Jens.Vigen@cern.ch).

Indexed in the CERN Document Server and in INSPIRE.

Published Open Access to permit its wide dissemination, as knowledge transfer is an integral part of the mission of CERN.

Proceedings of the 2023 European School on High-Energy Physics

Editors: Sascha Stahl and Alexander Huss

Abstract

The European School of High-Energy Physics is intended to give young physicists an introduction to the theoretical and experimental aspects of recent advances in elementary particle physics. These proceedings contain lecture notes on the theory of quantum chromodynamics, neutrino physics, flavour physics, cosmology, machine learning and practical statistics for particle physics.

Keywords

Higgs physics, Neutrino physics, QCD, Cosmology, Flavour physics, Statistics, Machine learning, Dark matter, Heavy ion physics, Standard Model, Lecture notes

Preface

The twenty-ninth event in the series of the European School of High-Energy Physics took place in Grenaa, Denmark, from 6 to 19 September 2023. It was organized by CERN, with support from Aarhus University, Copenhagen University, University of Southern Denmark and NICE: National Instrument Center for CERN Experiments. The local organization team was chaired by Stefania Xella (Copenhagen University). A total of 93 students of 37 different nationalities attended the school, mainly from institutes in member states of CERN, but also some from other regions. The participants were generally students in experimental High-Energy Physics in the final years of work towards their PhDs.

The School was hosted at the Kyst Hotellet in Grenaa (about 1 hour from Aarhus). According to the tradition of the School, the students shared twin rooms mixing participants of different nationalities.

A total of 31 lectures were complemented by daily discussion sessions led by six discussion leaders. The students displayed their own research work in the form of posters in an evening session in the first week, and the posters stayed on display until the end of the School. The full scientific programme was arranged in the on-site conference facilities.

The School also included an element of outreach training, complementing the main scientific programme. This consisted of a two-part course from the Inside Edge media training company. Additionally, students had the opportunity to act out radio interviews under realistic conditions based on a hypothetical scenario. The students from each discussion group subsequently carried out a collaborative project, preparing a talk on a physics-related topic at a level appropriate for a general audience. The talks were given by student representatives of each group in an evening session in the second week of the School. A jury, chaired by Mads Frandsen (University of Southern Denmark), judged the presentations; other members of the jury were Pamela Ferrari (Nikhef), Christophe Grojean (DESY) and Kate Ross (CERN). We are very grateful to all of these people for their help.

Our thanks go to the local-organization team for all of 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 efficient and friendly 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, appreciated listening to and discussing with the teaching staff of world renown. In addition to the rich academic programme, the participants enjoyed leisure and cultural activities in Denmark. There was a half-day excursion to Molsbjerge National Park and Ebeltoft, a traditional Danish town. A full-day excursion was organised to Aarhus, stopping first at the interesting Moesgaard Museum and then free time to explore the city of Aarhus. On the final Saturday afternoon, the students were able to make use of the hotel facilities and the local beach during free time. The excursions provided an excellent environment for informal interactions between staff and students.

We are very grateful to the School Administrator, Kate Ross (CERN), for her untiring efforts in the lengthy preparations for and the day-to-day operation of the School. Her continuous care of the participants and their needs during the School was highly appreciated.

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

Sascha Stahl[a]
(On behalf of the Organizing Committee)
[a]CERN

Lecture summaries

Introduction to perturbative QCD

The aim of these lectures is to give a brief introduction to perturbative QCD; focusing on basic concepts, fixed-order calculations, and phenomenological aspects.

Cosmology

This series of lectures covers the basics of cosmology from a particle physics point of view. The following topics will be partially covered: expanding Universe, cosmological parameters, generic approach to physical processes in the early Universe, cosmic microwave background radiation, nucleosynthesis, baryogenesis, dark matter, cosmological phase transitions and inflation.

Practical statistics excerpts from 2023 European School of High Energy Physics

These lecture notes provide an overview of fundamental statistical concepts used in high-energy physics research. The discussion begins with the philosophy of statistics, emphasizing its role in analyzing experimental data, correcting biases, and ensuring precision. Key estimators such as mean, standard deviation, skewness, and kurtosis, which help summarize datasets, are introduced. Various probability density functions (PDFs), including Binomial, Poisson, and Gaussian distributions, are explored to demonstrate their relevance in data modeling. Methods for analyzing data, such as maximum likelihood estimation and ChiSquare tests, are presented, offering techniques for extracting meaningful insights from observations. The notes also cover hypothesis testing, explaining concepts like false positive rates, p-values, and significance levels for evaluating scientific claims. Practical applications include setting observational limits and improving statistical methodology. By applying these techniques, researchers can ensure rigorous data analysis, enabling reliable conclusions and impactful discoveries in physics.

Neutrino physics

Surpassing all expectations, the Standard Model has predicted the outcomes of nearly every experiment conducted thus far. Neutrinos have no mass in it. However, we have gathered strong evidence in the last twenty years suggesting that the neutrinos have small but non-zero masses, These masses are a real delight because they allow neutrinos to oscillate and change flavor. I go over the characteristics of neutrinos both inside and outside the Standard Model in these lectures, as well as their incredible potential. I also revise the bits of data that defy the standard picture of the three neutrinos and discuss the possibilities of employing neutrinos to uncover any physics hidden outside the Standard Model.

Flavour physics and CP violation

We give a brief introduction into quark flavour physics and CP violation, starting in the first lecture with a review of the fundamental properties of the Standard Model of Particle Physics, a detailed discussion of the CKM matrix and a general classification of hadronic weak decays. The second lecture is devoted to describing the theoretical framework and in particular the concept of an effective Hamiltonian. In the third lecture we discuss mixing of neutral mesons and the effect of CP violation in hadron decays.

Machine learning

Advanced machine learning techniques have become ubiquitous: from computer vision algorithms found on a plethora of small devices such as cameras or smartphones to the recent rise of tremendously powerful large language models. Also in high energy particle physics, these techniques have become essential and have led to a significant increase in physics reach, from simple feed-forward algorithms used to distinguish signal and background processes to more complex neural networks that utilise the underlying physics structure of the data. This section will cover the basics of neural networks and their training and will then discuss examples of the building blocks that make up modern machine learning algorithms, aiming to provide a tool box for their further application in physics analyses.

Heavy-ion physics

Collisions of heavy ions at collider energies provide us with a unique opportunity to study strongly interacting matter at extreme temperatures and densities in the laboratory. Under these conditions quarks and gluons become deconfined to form a new state of matter, the quark-gluon plasma. Heavy ion physics has seen three major discoveries in the last 30 years. The first is that the QGP is the least dissipative material known and behaves like an almost perfect liquid. The second is that jets which are the manifestations of highly energetic quarks and gluons are strongly suppressed and modified compared to proton–proton collisions. This so-called jet quenching can be understood as the partial equilibration of a far-from-equilibrium system in a thermal QGP. The third main discovery is that particles with low transverse momentum produced in small collision systems like high multiplicity proton–proton and proton–ion collisions show many features that were believed to be signs for QGP formation. On the other hand, no jet quenching has been observed so far in small collision systems. These lectures are meant to give an overview over all relevant aspects of heavy ion physics at a phenomenological level.

Introduction to perturbative QCD

Gudrun Heinrich[a]
[a]KIT, Karlsruhe, Germany

The aim of these lectures is to give a brief introduction to perturbative QCD; focusing on basic concepts, fixed-order calculations, and phenomenological aspects.

1 Motivation: the era of precision phenomenology

Since the spectacular completion of the Standard Model through the Higgs discovery in 2012, the Large Hadron Collider (LHC) has delivered large amounts of high quality data and will continue to do so. To date, no direct evidence for physics beyond the Standard Model has been found at the LHC; however, for new physics that resides at energy scales beyond the LHC reach, small deviations from the Standard

This chapter should be cited as: Introduction to perturbative QCD, Gudrun Heinrich, DOI: 10.23730/CYRSP-2025-009.7,
in: Proceedings of the 2023 European School of High-Energy Physics,
CERN Yellow Reports: School Proceedings, CERN-2025-009, DOI: 10.23730/CYRSP-2025-009, p.7.

Model rather than the production of new particles will be the signposts. Therefore, high precision for both, theory predictions and measurements, are the name of the game.

In perturbative quantum field theory, precision is closely related to the calculation of higher orders in an expansion in the strong coupling α_s and the electroweak coupling α. For example, a cross section can be expanded in a power series in α_s as

$$\sigma = \sigma^{\rm LO} + \alpha_s \sigma^{\rm NLO} + \alpha_s^2 \sigma^{\rm NNLO} + \dots \;, \tag{1}$$

where LO means "leading order", NLO "next-to-leading order", NNLO "next-next-to-leading order", referring to the order of the perturbative expansion in the coupling. The leading-order cross section σ^{LO} itself may contain nonzero powers of α_s or α already, while the above power series indicates the QCD corrections, which, at NLO, can be classified into *real* and *virtual* corrections. The real corrections implies the radiation of extra QCD partons (gluons or quarks), while the virtual corrections contain loops of extra virtual particles with QCD couplings. We will go through explicit examples in this lecture.

As the value of the strong coupling at the energy scale of the Z-boson mass is $\alpha_s(M_Z) \simeq 0.118$, while the electroweak (EW) coupling, $\alpha(M_Z)$, amounts to about $1/128$, QCD corrections of the same power in the coupling are in general larger than EW corrections. As a rule of thumb, the NLO QCD corrections are typically $\mathcal{O}(10\%)$, NNLO QCD corrections a few %, and NLO EW corrections also a few %. However, there are important cases where this rule does not apply, for example

- Higgs production in gluon fusion: the NLO QCD corrections are $\mathcal{O}(100\%)$, and the NNLO corrections are still in the 40% range, see Fig. 1;
- kinematic regions where EW corrections are enhanced due to large logarithms of the form $\ln\left(\frac{M_V^2}{\hat{s}}\right)$, where $\hat{s} \gg M_V^2$ and M_V is the mass of a weak boson;
- kinematic regions dominated by soft and/or collinear radiation, where large logarithms of the form $\ln\left(\frac{M^2}{p_\perp^2}\right)$ occur, for example in the transverse momentum spectrum of a particle or particle pair with invariant mass M^2 at low transverse momentum $p_\perp$. As the infrared behaviour of QED and perturbative QCD is universal, i.e. process independent, such logarithmic terms can be predicted and the perturbative series can be re-organised. This is called *resummation*;
- mixed QCD $\otimes$ EW corrections also need to be considered when percent level precision is aimed at.

At the upcoming runs of the LHC, in particular the High-Luminosity LHC, percent level statistical uncertainties will be reached for key processes. Therefore, control of the theoretical uncertainties at this level is mandatory.

2 Basics of QCD

Quantum Chromodynamics (QCD) is the theory of the strong interactions between quarks and gluons. The emergence of QCD from the quark model [38, 43, 78] started more than 50 years ago, for a review see e.g. Ref. [48]. QCD as the theory of strong interactions is nowadays well established, however it still gives us many puzzles to solve and many tasks to accomplish in order to model particle interactions in collider physics to very high accuracy.

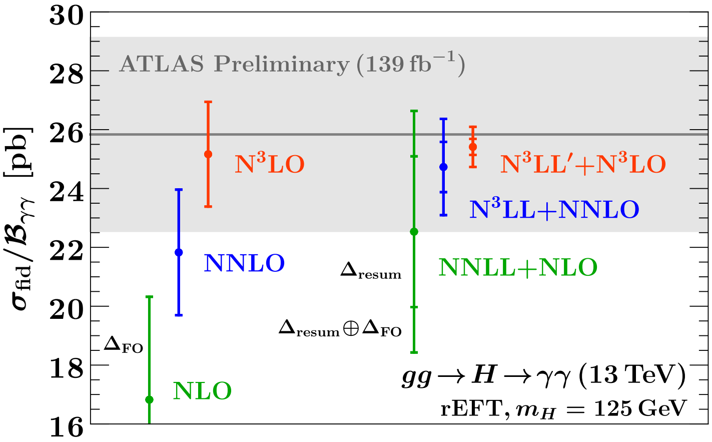


Fig. 1: Higher order QCD corrections for the fiducial cross section $gg \to H \to \gamma\gamma$, including q_T-resummation. Figure from Ref. [11].

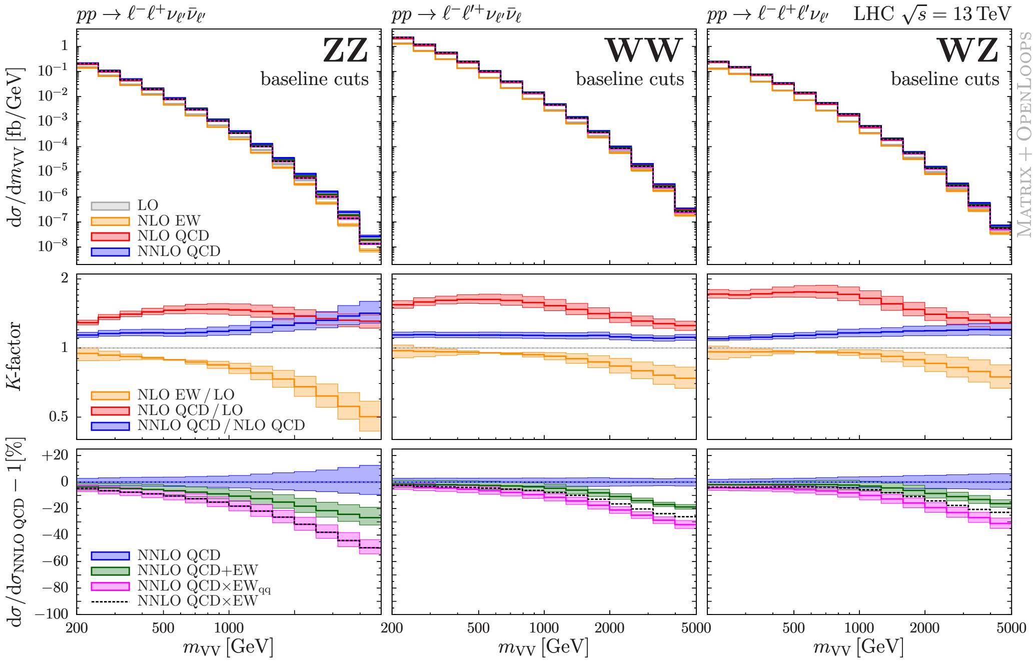


Fig. 2: Effect of electroweak corrections on vector boson pair production. Figure from Ref. [45].

The interactions are called "strong" since they are the strongest among the four fundamental forces at a length scale a bit larger than the proton radius. At a distance of 1 fm (1 fm = 10^{-15}m, the proton radius is about 0.88 fm), its strength is about 10^{38} times larger than the gravitational force. However, we will see later that the strong coupling is not constant, it varies with energy. The higher the energy at which we probe the interaction (i.e. the smaller the distance between the partons,[1] the weaker it will be. This phenomenon is called *asymptotic freedom*. At large distances between the partons, however, the interaction (coupling) becomes very strong. Therefore quarks and gluons cannot be observed as isolated particles. They are *confined* within hadrons, which are bound states of several partons.

[1] We will work in units where $\hbar = c = 1$. The mass of a proton in these units is $m_p \simeq 1\,\text{GeV} = 10^9$ eV. The conversion factor is $\hbar\, c \simeq 197.33$ MeV× fm.

Quarks come in six different *flavours*, denoted by u, d, c, s, t, b (up, down, charm, strange, top, bottom), grouped into three families. Hadrons can further be classified into baryons and mesons. Baryons are bound states of three quarks, for example the proton *(uud)*, mesons are quark–antiquark bound states. Pentaquarks also have been observed to exist as intermediate states.

Why "Chromodynamics"? In addition to the well-known quantum numbers like electromagnetic charge, spin or parity, quarks carry an additional quantum number called *colour*. Bound states are colour singlets. Note that without the colour quantum number, a bound state consisting e.g. of 3 u-quarks (called Δ^{++}) would violate Pauli's exclusion principle if there was no additional quantum number.

There are various approaches to make predictions and simulations based on QCD. They can be put into two broad categories: (i) perturbative QCD, which requires the coupling to be small, (ii) non-perturbative QCD, where QCD is in the strong coupling regime, such that perturbation theory is not applicable and methods such as for example "Lattice QCD" are appropriate. Our subject will be perturbative QCD.

2.1 Factorisation

A typical event at a hadron collider like the LHC is quite complicated, as sketched in Fig. 3. Since the proton is a bound state of quarks and gluons, non-perturbative aspects play a role in the initial state, and of course also in the final state, because quarks and gluons hadronize. Only the partonic cross section $\hat{\sigma}_0$ and the parton branchings can be described perturbatively. It is quite non-trivial that we can describe such complex interactions with high accuracy. The fact that we can separate the event into a part that is calculable in perturbation theory, the *hard scattering cross section*, and a non-perturbative part, is called *factorisation*. Factorisation would not be possible without *asymptotic freedom*, the fact that the strong coupling $\alpha_s(Q^2)$ decreases as the energy scale Q^2 increases.

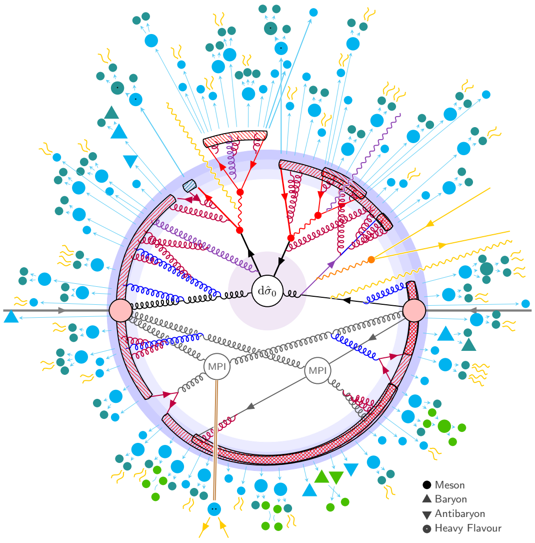


Fig. 3: Schematic picture of a LHC event. *Figure by T. Sjøstrand [48].*

The basic formula for factorisation at hadron colliders is the convolution of parton distribution functions (PDFs) with the partonic cross section.

For example, the differential cross section for a process like the production of a Higgs boson, $p_a + p_b \to H + X$, has the form

$$d\sigma_{pp\to H+X} = \sum_{i,j} \int_0^1 dx_1\, f_{i/p_a}(x_1, \alpha_s, \mu_f) \int_0^1 dx_2\, f_{j/p_b}(x_2, \alpha_s, \mu_f) \\ \times d\hat{\sigma}_{ij}(x_1, x_2, \alpha_s(\mu_r), \mu_r, \mu_f) \;+\; \mathcal{O}\left(\frac{\Lambda}{Q}\right)^p . \tag{2}$$

The hard scattering cross section $d\hat{\sigma}_{ij}$ is factorised from the non-perturbative parton distribution functions f_{i/p_a}, f_{j/p_b} at the factorisation scale μ_f. Factorisation, shown schematically in Fig. 4, holds up to the so-called *power corrections* of order $\left(\frac{\Lambda}{Q}\right)^p$. The power p depends on the observable, but is always positive, and $\Lambda \simeq 250\,\mathrm{MeV}$ is the scale where non-perturbative effects start to dominate the behaviour of the strong coupling. Therefore, the larger the typical energy scale Q^2 of the hard process, the smaller the power corrections.

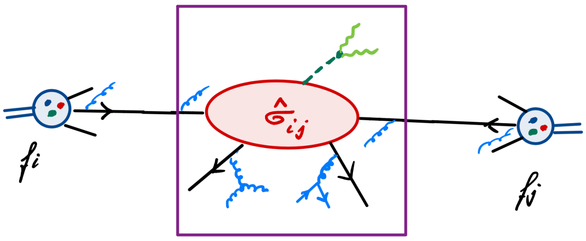


Fig. 4: Schematic picture of factorisation in hadron–hadron collisions. The purple box denotes the part that can be calculated perturbatively.

Strictly speaking, the above factorisation is called “collinear factorisation”, because it assumes that the partons which are coming from the hadrons a and b have a momentum which is collinear to the parent hadron momentum. There is also the so-called “transverse momentum dependent (TMD)” factorisation, which takes into account that the partons inside the proton can have a transverse momentum relative to the beam axis. This requires transverse momentum dependent PDFs, see e.g. Ref. [70] for a review. The TMD effects can become sizeable for example in Drell-Yan production ($pp \to V \to l^+l^-$) at very low transverse momenta of the produced vector boson, or in semi-inclusive DIS (deep-inelastic scattering).

We will discuss the PDFs as well as asymptotic freedom in more detail later, here we want to stress that without factorisation, it would not be possible to produce the high precision predictions as we have them today within a strongly interacting theory, i.e. QCD.

As the PDFs themselves cannot be calculated from first principles, but need to be fitted from data, their contribution to the uncertainty budget increases in relative size the more higher orders in perturbation theory are available for the partonic cross section. The situation as of 2018 for the case of Higgs production in gluon fusion is shown in Fig. 5. While some of these uncertainties, such as $\delta(1/m_t)$, $\delta(t,b,c)$ and $\delta(EW)$, have been reduced substantially in the past few years [8, 13, 24, 25, 27], the PDF-related uncertainties are still an issue.

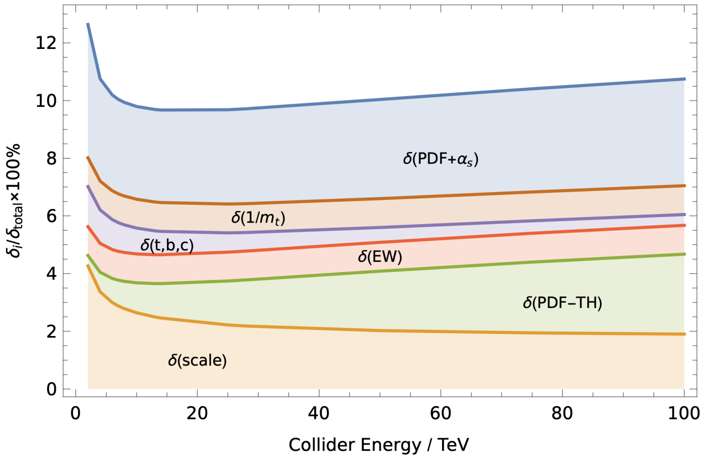


Fig. 5: Contributions to the total uncertainty for Higgs production in gluon fusion. Figure from Ref. [35].

2.2 QCD Lagrangian and Feynman rules

2.2.1 Colour algebra

The strong interactions can be described as an $SU(3)$ local gauge theory, where the "charges" are denoted as *colour*, therefore the name "Quantum Chromodynamics" (QCD). The strong interactions are embedded in the Standard Model with underlying gauge group structure $SU(3) \times SU(2)_L \times U(1)_Y$. Often the number of colours in denoted generically by N_c and the colour algebra in QCD calculations is done for a generic $SU(N_c)$ gauge group. Experimental evidence suggests that in nature $N_c = 3$, but the concept is more general. Using a generic number of colours N_c has many advantages, for example it allows us to divide the amplitudes into simpler building blocks according to their colour structure.

The group $SU(N_c)$ is an example of so-called *Lie groups* (named after Sophus Lie, Norwegian mathematician, 1842-1899), which are briefly discussed below. As we will see, the non-Abelian group structure of $SU(N_c)$ implies that gluons interact with themselves (while photons do not).

Groups

A group G is a set of elements g with a multiplication law "$\circ$": $G \times G \to G$ which satisfies:

- there is a unit element e with $g \circ e = e \circ g = g$,
- for each g an inverse g^{-1} exists, with $g \circ g^{-1} = g^{-1} \circ g = e$,
- associativity: $g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3$.

For *Abelian* groups, in addition comutativity, $g_1 \circ g_2 = g_2 \circ g_1$, holds (named after Niels Henrik Abel, Norwegian mathematician, 1802-1829).

We will deal with compact Lie groups, which are groups whose elements depend analytically on a finite number of continuous parameters. The group $SU(N)$ is a Lie group whose representations U are unitary matrices with determinant one, $UU^\dagger = 1 \wedge \det U = 1$.
Other examples of Lie groups are the orthogonal groups $SO(N)$, e.g. the rotation group $SO(3)$, the symplectic groups $Sp(N)$ or the special groups G_2, F_4, E_6, E_7, E_8.

Representations

A *representation* of a group is a mapping of the group elements onto matrices, where the group multiplication laws translate to matrix multiplication (preserving the group multiplication laws), i.e. linear algebra can be used.

For Lie groups, any group element can be obtained from the identity by continuous changes in the parameters, i.e. it can be written as

$$U = \exp\{i\,\theta_a T^a\}\ ,\ \theta_a \in \mathbb{R}\ . \tag{3}$$

From $UU^\dagger = 1$ it follows that $T^a = (T^a)^\dagger$ (hermitian), from $\det U = 1$ we derive $\mathrm{Tr}(T^a) = 0$ (using $\det U = e^{\mathrm{Tr}(\ln U)}$). The set of all linear combinations $T^a\theta^a$ is a vector space and the T^a form a basis in that space. Therefore they are also called the *generators* of the group.
For the generators T^a, the commutation relation

$$[T^a,\, T^b] = i\, f^{abc}\, T^c \tag{4}$$

holds, independent of the representation, defining an algebra associated with the group. The f^{abc} are called *structure constants*. For Abelian groups the structure constants are zero.

Further, the generators satisfy the *Jacobi identity*

$$[T^a, [T^b,\, T^c]] + [T^b, [T^c,\, T^a]] + [T^c, [T^a,\, T^b]] = 0\ , \tag{5}$$

wich translates into a relation between structure constants. The generators are normalised such that

$$\mathrm{Trace}(T^a T^b) = T_R\, \delta^{ab}\, . \tag{6}$$

Usually one chooses $T_R = 1/2$ for the fundamental representation, $R = F$, which we will also do. While the (totally antisymmetric) structure constants are given in terms of generators by

$$f^{abc} = -2i\,\mathrm{Trace}\Big(T^a\big[T^b, T^c\big]\Big)\ , \tag{7}$$

we can also define totally symmetric constants by

$$d^{abc} = 2\,\mathrm{Trace}\Big(T^a\big\{T^b, T^c\big\}\Big)\ . \tag{8}$$

For QCD, two representations of $SU(N)$ will be important:
1. the *fundamental representation*: the generators are $N \times N$ matrices,
2. the *adjoint representation*: the generators of this representation are $(N^2 - 1) \times (N^2 - 1)$-matrices, i.e. the indices run over the same range as the number of generators. The number of generators is called the *dimension* of the group. So in the adjoint representation, the dimension of the vector space in which the representation matrices act equals the dimension of the group. Therefore the generators in the adjoint

representation can be expressed in terms of structure constants:

$$T^a_{bc} \underset{\text{adj}}{=} (F^a)_{bc} =: -i\, f^{abc}\ ,\ a, b, c = 1 \ldots N^2 - 1\ . \tag{9}$$

The generators of $SU(3)$ in the fundamental representation are usually defined as $t^a_{ij} = \lambda^a_{ij}/2$, where the λ^a_{ij} are also called *Gell–Mann* matrices. They are traceless and hermitian and can be considered as the $SU(3)$ analogues of the Pauli matrices for $SU(2)$.

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

$$\lambda^4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}\ ,\lambda^5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}\ ,\ \lambda^6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}\ ,$$

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

Colour in QCD

Quark fields transform according to the fundamental representation of $SU(3)$. Therefore the Feynman rules for the quark–gluon vertex involve t^a_{ij} where $i, j = 1 \ldots N_c$ run over the colours of the quarks (the *degree* of the group), while $a = 1 \ldots N_c^2 - 1$ runs over the dimension of the group. Gluons transform according to the adjoint representation of $SU(3)$. Therefore the Feynman rule for the three-gluon vertex contains $(F^a)_{bc} = -i\, f^{abc}$.

The gluons can be regarded as a combination of coloured lines, as depicted in Fig. 6, or, more precisely, as a combination of colours and anticolours. Contracting colour indices is graphically equivalent to connecting the respective colour (or anticolour) lines. The representation of the quark–gluon vertex in Fig. 6 embodies colour conservation, whereby the colour–anticolour quantum numbers carried by the $q\bar{q}$ pair are transferred to the gluon. The eight gluons can be regarded as all possible colour–anticolour

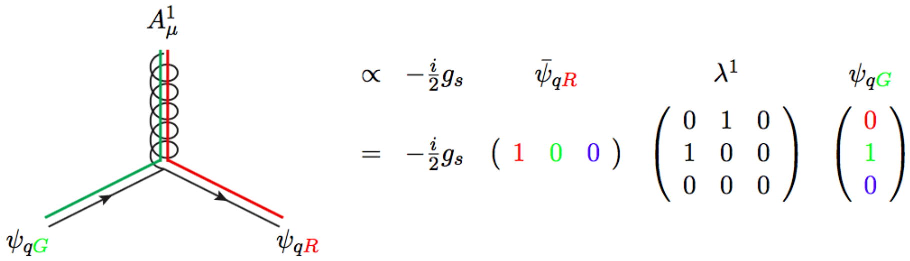

A^1_μ
ψ_{qG}
ψ_{qR}
$\propto\ -\frac{i}{2}g_s \quad \bar{\psi}_{qR} \quad \lambda^1 \quad \psi_{qG}$
$= -\frac{i}{2}g_s \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

Fig. 6: Representation of the gluon as a double colour line. *Picture from Ref. [72].*

combinations, where the combination corresponding to λ^8 would be $\frac{1}{\sqrt{6}}(r\bar{r} + g\bar{g} - 2b\bar{b})$ (with r: red, g: green, b: blue). Note that the singlet combination $\frac{1}{\sqrt{3}}(r\bar{r} + g\bar{g} + b\bar{b})$ does not occur for the gluon because the singlet cannot mediate colour.

Casimir operators

The sums $\sum_{a,j} t^a_{ij} t^a_{jk}$ and $\sum_{a,d} F^a_{bd} F^a_{dc}$ have two free indices in the fundamental and adjoint representation, respectively. One can show that these sums are invariant under $SU(N_c)$ transformations, and therefore must be proportional to the unit matrix:

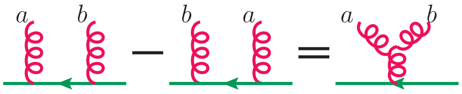
$$\sum_{j,a} t^a_{ij} t^a_{jk} = C_F\, \delta_{ik}\,, \qquad \sum_{a,d} F^a_{bd} F^a_{dc} = C_A\, \delta_{bc}\,. \tag{10}$$

The constants C_F and C_A are the eigenvalues of the quadratic *Casimir operator* in the fundamental and adjoint representation, respectively. Casimir operators commute with every element of the group. The Casimir eigenvalues can be expressed in terms of the number of colours N_c as

$$C_F = T_R \frac{N_c^2 - 1}{N_c}\ , \ C_A = 2\, T_R\, N_c\ . \tag{11}$$

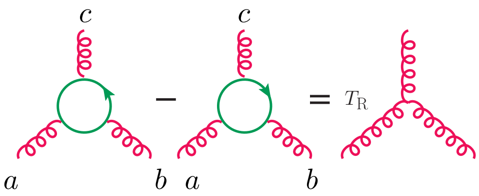
Graphical representation of the colour algebra

The commutation relation (4) in the fundamental representation can be represented graphically by[2]



$$t^a t^b - t^b t^a = i\, f^{abc}\, t^c$$

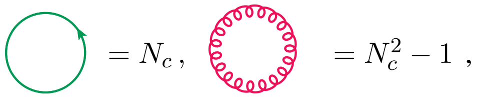
Multiplying this commutator first with another colour charge operator, summing over the fermion index and then taking the trace over the fermion line (i.e. multiplying with δ_{ik}) we obtain the representation of the three-gluon vertex as traces of products of colour charges:



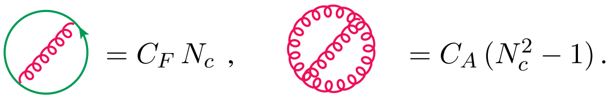
$$\text{Trace}(t^a t^b t^c) - \text{Trace}(t^c t^b t^a) = i\ T_R f^{abc} \tag{12}$$

Such relations allow us to compute the colour algebra structure of a QCD diagram, independent of the kinematics. For example, taking the trace of the identity (δ_{ij} resp. δ_{ab}) in the fundamental resp. in the adjoint representation we obtain

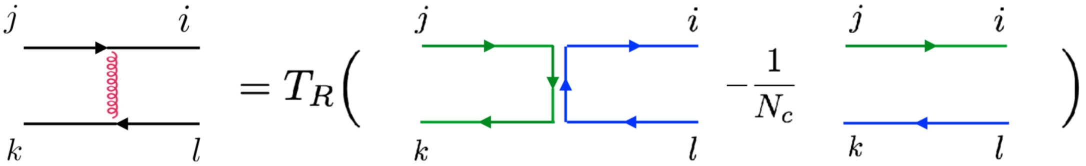
[2] Some of the figures have been taken from Ref. [75].



respectively. Then, using the expressions for the fermion and gluon propagator insertions, we find



There is also a very useful identity for the product of two colour matrices in the fundamental representation, occurring when a gluon is exchanged between two quark lines, and following from representing the gluon as a double quark line,



corresponding to

$$t^a_{ij} t^a_{kl} = T_R \left(\delta_{il}\delta_{kj} - \frac{1}{N_c}\delta_{ij}\delta_{kl} \right) \ . \tag{13}$$

The second term $\sim 1/N_c$ implements the condition that the generators are traceless, and the picture indicates that a gluon which mediates between quarks of the same colour does not exist, because it would be a colour singlet.

Colour decomposition

From the representation of gluons as double colour lines it follows that any tree-level diagram for n-gluon scattering can be expressed in terms of traces over generators t^a_{ij}, as depicted in Fig. 7. This observation leads to the so-called *colour decomposition* of amplitudes, which allows to separate the colour information from the kinematic part of an amplitude. An amplitude for n-gluon scattering can be

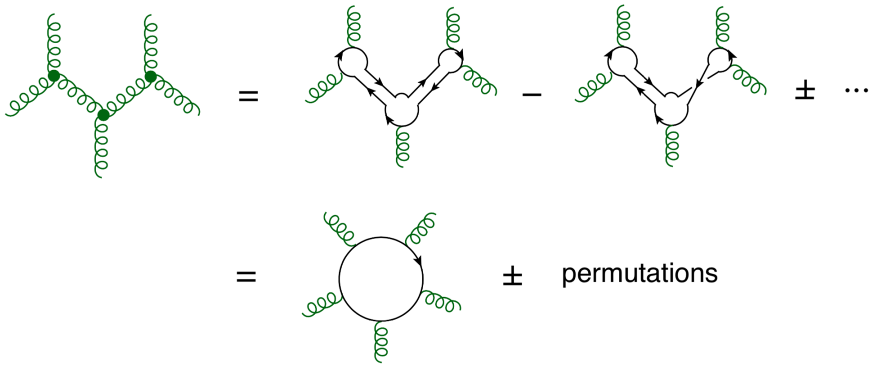


Fig. 7: Colour decomposition of tree-level gluon amplitudes. *Figure from Ref. [31].*

written as

$$\mathcal{A}_n^{\text{tree}}\left(\{k_i,\lambda_i,a_i\}\right) = g_s^{n-2} \sum_{\sigma\in S_n/Z_n} \text{Trace}\left(t^{a_{\sigma(1)}}\cdots t^{a_{\sigma(n)}}\right) A_n^{\text{tree}}(\sigma(1^{\lambda_1}),\ldots,\sigma(n^{\lambda_n}))\;, \tag{14}$$

where k_i, λ_i are the gluon momenta and helicities, a_i the colour indices and $A_n^{\text{tree}}(1^{\lambda_1},\ldots,n^{\lambda_n})$ are the *partial amplitudes*, which contain all the kinematic information. S_n is the set of all permutations of n objects, while Z_n is the subset of cyclic permutations, which preserves the trace; the latter are excluded in the sum over the set S_n/Z_n.

The advantage of this representation is that the partial amplitudes A_n^{tree} are simpler to calculate than the full amplitude because they are *colour ordered*: they only receive contributions from diagrams with a particular cyclic ordering of the gluons. This implies that for the partial amplitudes, the infrared singularities related to external massless particles becoming collinear can only occur in a subset of momentum channels, those with cyclically adjacent momenta. For example, the five-point partial amplitudes $A_5^{\text{tree}}(1^{\lambda_1},2^{\lambda_2},3^{\lambda_3},4^{\lambda_4},5^{\lambda_5})$ can only have poles in s_{12}, s_{23}, s_{34}, s_{45}, s_{51}, and not in s_{13}, s_{24}, s_{35}, s_{41} or s_{52}, where $s_{ij}\equiv(k_i+k_j)^2$.

The colour decomposition is not limited to gluons only, it also can be applied when quarks are involved. For example, a tree amplitude with a $q\bar{q}$ pair and otherwise gluons can be written as

$$\mathcal{A}_n^{\text{tree}} = g_s^{n-2} \sum_{\sigma\in S_{n-2}} \left(t^{a_{\sigma(3)}}\cdots t^{a_{\sigma(n)}}\right)_{j_1 i_2} A_n^{\text{tree}}(1_{\bar{q}}^{\lambda_1},2_q^{\lambda_2},\sigma(3^{\lambda_3}),\ldots,\sigma(n^{\lambda_n}))\;. \tag{15}$$

Using eq. (13), there is yet another possibility to perform the colour decomposition, based on Kronecker δ_{ij}'s only, also called *colour flow decomposition* [62]. The computational gain when using colour decomposition to calculate amplitudes with n gluons is illustrated in Table 1. At loop level, colour decomposition can also be performed [9,29,31,53,65,68].

	# diagrams	
n	partial amplitude	full amplitude
4	3	4
5	10	25
6	36	220
7	133	2485
8	501	34300
9	1991	559405
10	7335	10525900
11	28199	224449225
12	108281	5348843500

Table 1: Number of diagrams for tree-level n-gluon amplitudes. *Table from Ref. [62].*

Another advantage of colour decomposition is the possibility to approximate complex calculations by the leading-colour approximation, where only the leading term in N_c is retained. A detailed study of subleading colour effects in single jet inclusive and dijet production at NNLO can be found in Ref. [20].

2.2.2 *Experimental evidence for the existence of colour*

The colour factors C_F and C_A can be measured indirectly from jet production cross sections. Jets can be pictured as clusters of particles (usually hadrons) which are close to each other in phase space, resp. in the detector, see Fig. 8, showing a 6-jet event measured by CMS. Jets will be discussed in more detail in Section 4.

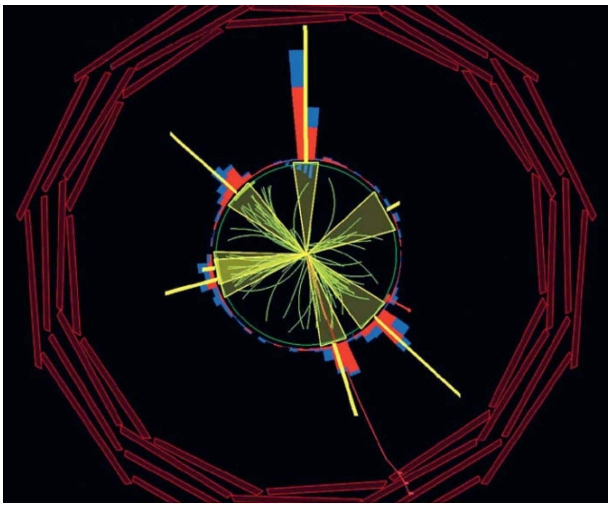


Fig. 8: Example of a 6-jet event measured by CMS. *Source: CERN Courier*.

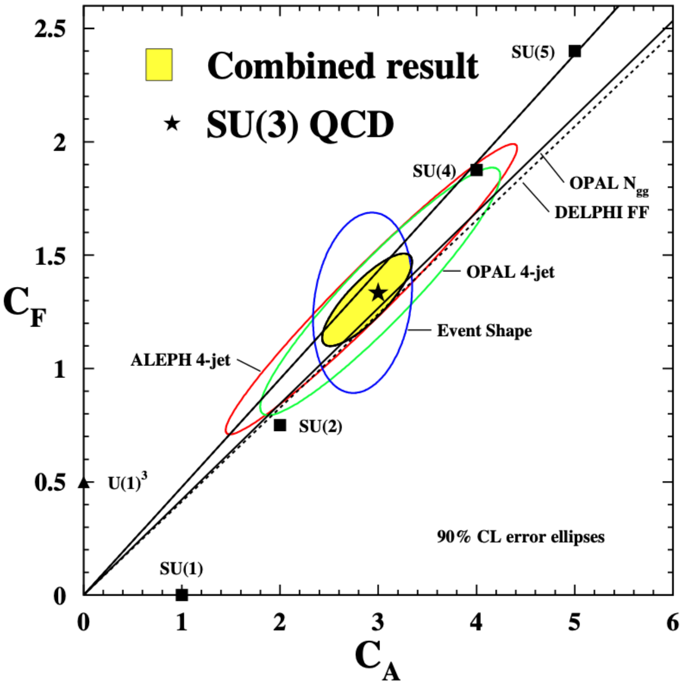


Fig. 9: Evidence for colour-$SU(3)$ from jet measurements at the LEP collider. Figure from Ref. [56].

The theory predictions for the jet cross sections depend on the number of colours, as can be seen in Fig. 9 for a measurement of jet cross sections at LEP. The leading-colour contribution to the squared

n-gluon matrix elements relevant at hadron colliders, summed over colours (and helicities), is given by

$$|\mathcal{M}_n|^2 = \left(g^2 N_c\right)^{n-2} (N_c^2 - 1) \sum_{\sigma \in S_n/Z_n} \left\{ |\mathcal{A}_n(\sigma(1), \cdots, \sigma(n))|^2 + \mathcal{O}\left(\frac{1}{N_c^2}\right) \right\} ,$$

so if N_c was different, the jet cross sections would change drastically. How well they agree with recent measurements is shown in Fig. 10.

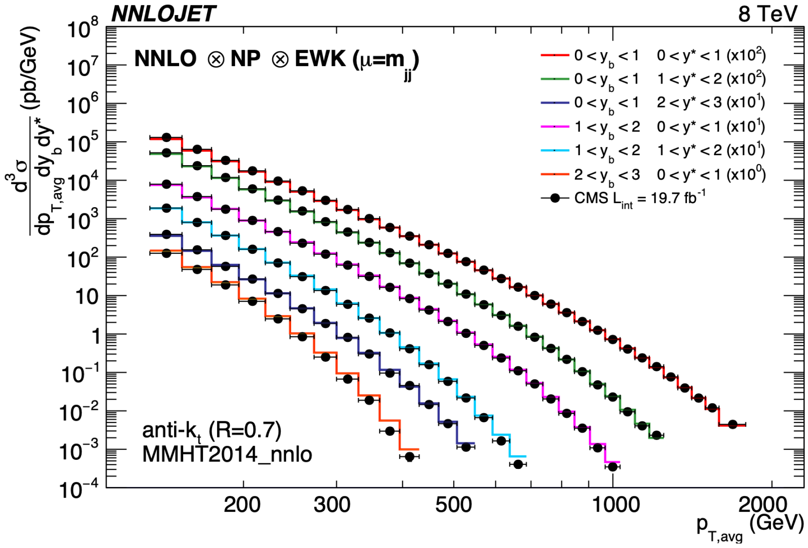


Fig. 10: Triple-differential two-jet cross sections at NNLO compared to CMS data. Figure from Ref. [42].

A good theoretical description of jets at the LHC is important for many reasons. For example, jet data are very important to constrain the PDFs and jet cross sections are used for precision determinations of the strong coupling α_s. Furthermore, for jets originating from massless partons, there is energy available to go up to very high values in the jet p_T spectrum, $\mathcal{O}$(2–4 TeV), which is a kinematic region sensitive to new physics. Jets recoiling against missing transverse energy (“missing” in the total transverse energy budget) could be a signal for an unknown heavy particle (for example related to dark matter) decaying into something not visible in the detector.

Hadronic R-ratio

Among the historically early evidence for the existence of 3 colour quantum numbers is the so-called *hadronic R-Ratio*, the total cross section for the production of hadrons in electron–positron collisions, divided by the cross section for the production of a muon–antimuon pair, as a function of the centre-of-mass energy s,

$$R(s) = \frac{\sigma(e^+e^- \to \text{ hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)} \; . \tag{16}$$

Hadrons are bound states of quarks and gluons, so for the numerator, at microscopic level, this means that a fermion–antifermion pair $f\bar{f}$ is created as soon as the centre-of-mass energy s is sufficient to produce two quarks of mass m_f . The interaction proceeds via the exchange of a virtual photon, see

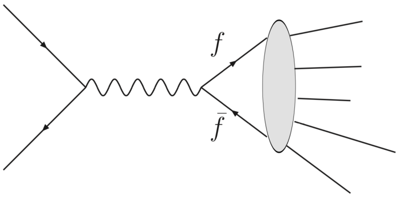


Fig. 11: The production of hadrons in e^+e^- collisions via virtual photon exchange.

Fig. 11 (neglecting Z-boson exchange). The electromagnetic interaction (the photon) only sees the electromagnetic charge e_f of the fermions, be it quarks or leptons. Therefore one would expect that at large energies

$$R(s) = \frac{\sum_{f=u,d,s,c,\ldots} \sigma(e^+e^- \to f\bar{f})}{\sigma(e^+e^- \to \mu^+\mu^-)} \overset{s \text{ large}}{\longrightarrow} \sum_{f=u,d,s,c,\ldots} e_f^2 \, \theta(s - 4m_f^2) \; . \tag{17}$$

However this is not what has been found experimentally. The experimental results agree with the expression

$$R(s) = N_c \sum_{f=u,d,s,c,\ldots} e_f^2 \, \theta(s - 4m_f^2) \; , \tag{18}$$

with $N_c = 3$. Above the bottom quark pair production threshold, we have ($m_b \simeq 4.2$ GeV)

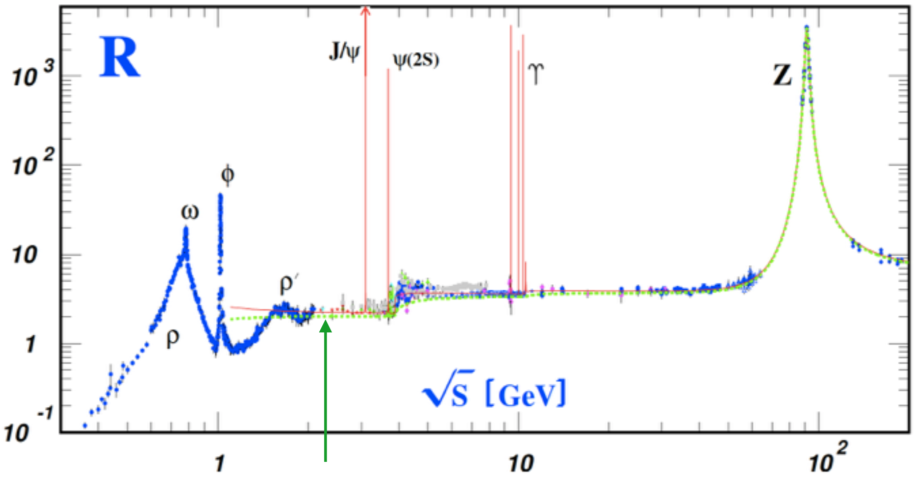


Fig. 12: R-Ratio vs. center-of-mass energy. *Figure from Ref. [57].*

$$R(s) = N_c \sum_{f=u,d,s,c,b} e_f^2 \, \theta(s - 4m_f^2) = 3\left(\tfrac{4}{9} + \tfrac{1}{9} + \tfrac{1}{9} + \tfrac{4}{9} + \tfrac{1}{9}\right) = \tfrac{11}{3} \; . \tag{19}$$

The stepwise increase of the R-ratio can be seen in Fig. 12, even though it is blurred by the resonances ρ, ω, ϕ, etc. In Fig. 13 the quark pair production thresholds without the resonances are shown. The top quark does not show up here, it is heavier than the Z-boson, $m_t \simeq 173$ GeV, and it decays before it hadronises.

Another example which is often given as an argument for $N_c = 3$ is the decay rate of a neutral

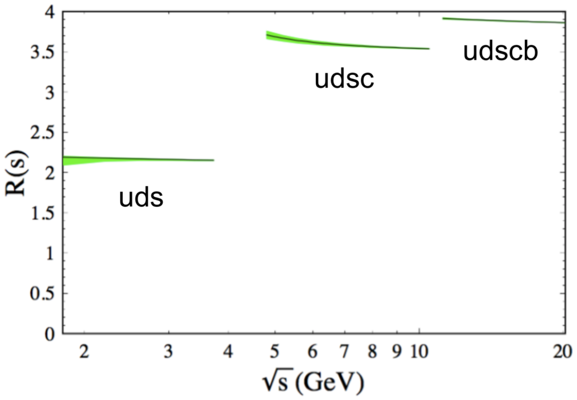


Fig. 13: R-ratio without resonance effects. *Figure from Ref. [50].*

.

pion into two photons:

$$\Gamma(\pi^0 \to \gamma\gamma) \simeq \alpha^2 \frac{m_\pi^3}{f_\pi^2}(e_u^2 - e_d^2)^2 N_c^2 \,. \tag{20}$$

However, there could be cancellations between N_c and a different denominator in the fractional charges of the quarks: for $e_u = (1/N_c + 1)/2$, $e_d = (1/N_c - 1)/2$ the decay rate would be independent of N_c (and a relation between N_c and the fractional charges of the quarks also makes sense in view of anomaly cancellation). Therefore, Eq. (20) alone cannot be seen as a full proof of N_c being equal to 3.

A more theoretical argument for the existence of a colour quantum number has been mentioned already: it is given by the fact that bound states consisting of three quarks, e.g. the delta-resonance Δ^{++} consisting of 3 u-quarks, would violate Pauli's exclusion principle if there was no additional quantum number (which implies that this state must be totally antisymmetric in the colour indices).

2.2.3 QCD Lagrangian

An important concept in QCD (and in quantum field theories in general) is the formulation as a *local* gauge theory. This means that the gauge transformation parameter depends itself on x, the position in space–time.

Fermionic part of the QCD Lagrangian

Consider the quark fields $q_f^j(x)$ for just one quark flavour f. The index j labels the colour, $j = 1, \dots, N_c$. Treating the quarks as free Dirac fields, we have

$$\mathcal{L}_q^{(0)}(q_f, m_f) = \sum_{j,k=1}^{N_c} \bar{q}_f^j(x)\,(i\,\gamma_\mu\partial^\mu - m_f)\delta_{jk}\,q_f^k(x)\,, \tag{21}$$

where the Dirac-matrices γ_μ satisfy the anti-commutation relation (Clifford algebra)

$$\{\gamma^\mu, \gamma^\nu\} \,=\, 2\,g^{\mu\nu}\,. \tag{22}$$

Now let us apply a group transformation (i.e. a "rotation" in colour space) on the fermion fields. It has the form

$$q_k \to q'_k = U_{kl}\, q^l \;, \quad \bar{q}_k \to \bar{q}'_k = \bar{q}^{\,l} U^{-1}_{lk} \;, \tag{23}$$

with

$$U_{kl} = \exp\left\{ i \sum_{a=1}^{N_c^2-1} t^a\, \theta^a \right\}_{kl} \equiv \exp\left\{ i\, \boldsymbol{t}\cdot\boldsymbol{\theta} \right\}_{kl} \;, \tag{24}$$

where θ^a are the group transformation parameters and $(t^a)_{kl}$ the generators of $SU(N_c)$ in the fundamental representation. The Lagrangian of free Dirac fields remains invariant under this transformation as long as it is a *global* transformation, i.e. as long as the θ^a do not depend on x: $\mathcal{L}_q^{(0)}(q) = \mathcal{L}_q^{(0)}(q')$.

However, we aim at *local* gauge transformations, where the gauge transformation parameter θ in Eq. (24) depends on x. In QED, where the underlying gauge group is $U(1)$, a global transformation would just be a phase change. The requirement of a free electron field to be invariant under *local* transformations $\theta = \theta(x)$ leads to the introduction of a *gauge field* A_μ, the photon. The analogous is true for QCD: requiring local gauge invariance under $SU(N_c)$ leads to the introduction of gluon fields A^a_μ to arrive at a gauge invariant Lagrangian, wich can be seen as follows:
As the local gauge transformation

$$U(x) = \exp\left\{ i\, \boldsymbol{t}\cdot\boldsymbol{\theta(x)} \right\} \tag{25}$$

depends on x, the derivative of the transformed quark field $q'(x)$ reads

$$\partial_\mu\, q'(x) = \partial_\mu \left(U(x) q(x) \right) = U(x)\partial_\mu\, q(x) + \left(\partial_\mu U(x) \right)\, q(x) \;. \tag{26}$$

To keep $\mathcal{L}_q$ gauge invariant, we can remedy the situation caused by the second term above by introducing the coupling to a gauge field which transforms accordingly. We define a *covariant derivative* D^μ, depending on A^μ_a, by

$$\left(D^\mu[A] \right)_{ij} = \delta_{ij}\partial^\mu + i\, g_s\, t^a_{ij} A^\mu_a \;, \tag{27}$$

or, without index notation

$$\boldsymbol{D}^\mu[\boldsymbol{A}] = \partial^\mu + i\, g_s\, \boldsymbol{A}^\mu \;, \tag{28}$$

where $\boldsymbol{A}^\mu = t^a\, A^\mu_a$ (sum over $a = 1 \ldots N_c^2 - 1$ understood). The fields A^μ_a are the *gluons*. The Lagrangian corresponding to this "minimal coupling" of a gluon field reads

$$\mathcal{L}_q(q_f,\, m_f) = \sum_{j,k=1}^{N_c} \bar{q}^{\,j}_f(x)\, \left(i\, \gamma_\mu \boldsymbol{D}^\mu[\boldsymbol{A}] - m_f \right)_{jk}\, q^k_f(x) \;. \tag{29}$$

To keep this Lagrangian invariant under local gauge transformations, we therefore must have

$$\begin{aligned} \boldsymbol{D}^\mu[\boldsymbol{A}']q'(x) &\overset{!}{=} U\Big(\boldsymbol{D}^\mu[\boldsymbol{A}]\, q(x) \Big) \\ \Rightarrow \partial_\mu + i g_s \boldsymbol{A}'_\mu &\overset{!}{=} U \left(\partial_\mu + i g_s \boldsymbol{A}_\mu \right) U^{-1} \end{aligned} \tag{30}$$

Eq. (30) gives a condition on $\boldsymbol{A}'_\mu(x)$, which can be derived using $U\,\partial_\mu U^{-1} = -(\partial_\mu U)U^{-1} + \partial_\mu(UU^{-1})$. Therefore the gluon fields need to transform under general $SU(N_c)$ transformations as

$$\boldsymbol{A}'_\mu = U(x)\boldsymbol{A}_\mu U^{-1}(x) + \frac{i}{g_s}(\partial_\mu U(x))U^{-1}(x)\ . \tag{31}$$

Purely gluonic part of the QCD Lagrangian

The purely gluonic part of the QCD Lagrangian can be described by the so-called Yang–Mills Lagrangian (C. N. Yang, R. Mills, 1954)

$$\mathcal{L}_{\mathrm{YM}} = -\frac{1}{4}F^a_{\mu\nu}\,F^{a,\mu\nu}\ , \tag{32}$$

where the non-Abelian field strength tensor $F^a_{\mu\nu}$ is given by

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g_s\,f^{abc}A^b_\mu A^c_\nu\ . \tag{33}$$

We can also express everything in terms of $\boldsymbol{A}^\mu = t^a\,A^\mu_a$ and write the field strength tensor as

$$\boldsymbol{F}_{\mu\nu}(x) = \sum_{a=1}^{N_c^2-1} F^a_{\mu\nu}(x)\,t^a = \frac{i}{g_s}\,[\,\boldsymbol{D}_\mu, \boldsymbol{D}_\nu\,]\ , \tag{34}$$

which implies

$$\mathcal{L}_{\mathrm{YM}} = -\frac{1}{4}\,F^a_{\mu\nu}\,F^{a,\mu\nu} = -\frac{1}{2}\,\mathrm{Trace}\,[\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu}]\ . \tag{35}$$

Note that the term proportional to f^{abc} in the expression for $F^a_{\mu\nu}$, which reflects the non-Abelian structure and is not present in QED, leads to terms with 3 or 4 gluon fields in the Lagrangian and therefore to self-interactions between the gluons.

Finally, we obtain for the “classical” QCD Lagrangian

$$\mathcal{L}_c = \mathcal{L}_{\mathrm{YM}} + \mathcal{L}_q = -\frac{1}{4}F^a_{\mu\nu}\,F^{a,\mu\nu} + \sum_{j,k=1}^{N_c} \bar{q}^j_f(x)\,\left(i\,\gamma_\mu D^\mu[A] - m_f\right)_{jk}\,q^k_f(x)\ . \tag{36}$$

Gauge fixing

We are not quite there yet with the complete QCD Lagrangian. The “classical” QCD Lagrangian $\mathcal{L}_c$ contains degenerate field configurations (i.e. they are equivalent up to gauge transformations). This leads to the fact that the bilinear operator in the gluon fields is not invertible, such that it is not possible to construct a propagator for the gluon fields. The propagator is usually derived from the bilinear term in the fields in the path integral for free fields, with the generating functional (introducing the d’Alembert operator $\Box \equiv \partial_\mu\partial^\mu$)

$$Z_0[J] = \int \mathcal{D}A_\mu(x)\,e^{i\int d^4x\left[\frac{1}{2}A^a_\mu(x)(g^{\mu\nu}\Box - \partial^\mu\partial^\nu)A^b_\nu(x)\delta_{ab} + J^a_\mu A^\mu_a\right]}\ . \tag{37}$$

In momentum space this leads to the following condition for the propagator $\Delta_{\mu\nu}(p)$, where we suppress colour indices as the propagator is diagonal in colour space, i.e. we leave out overall factors δ^{ab},

$$i\,\Delta_{\mu\rho}(p)\;\left[p^2 g^{\rho\,\nu}\;-\;p^\rho p^\nu\right] = g^\nu_\mu\,. \tag{38}$$

However, we also have

$$\left[p^2 g^{\rho\,\nu} - p^\rho p^\nu\right] p_\nu = 0\,, \tag{39}$$

which means that the matrix $\left[p^2 g^{\rho\,\nu}\;-\;p^\rho p^\nu\right]$ is not invertible because it has at least one eigenvalue equal to zero. We have to remove the physically equivalent configurations from the classical Lagrangian. This is called *gauge fixing*. We can achieve this by imposing a constraint on the fields A^a_μ, adding a term to the Lagrangian with a Lagrange multiplier.

For example, *covariant gauges* are defined by the requirement $\partial_\mu A^\mu\,(x) = 0$ for any x. Adding

$$\mathcal{L}_{\rm GF} = -\frac{1}{2\lambda}\,(\partial_\mu A^\mu)^2\,, \qquad \lambda\,\in\,\mathbb{R},$$

to $\mathcal{L}$, the bilinear term has the form

$$i\left(p^2 g^{\mu\nu} - \left(1-\frac{1}{\lambda}\right)p^\mu p^\nu\right)\,,$$

with inverse

$$\Delta_{\mu\nu}(p) = \frac{-i}{p^2+i\,\varepsilon}\left[g_{\mu\nu} - (1-\lambda)\frac{p_\mu p_\nu}{p^2}\right]\,. \tag{40}$$

The so-called $i\,\varepsilon$–prescription ($\varepsilon\,>\,0$) shifts the poles of the propagator slightly off the real p^0-axis (where p^0 is the energy component) and ensures the correct causal behaviour of the propagators. Choosing $\lambda = 1$ is called *Feynman gauge*, and $\lambda = 0$ is called *Landau gauge*. Of course, physical results must be independent of λ.

In covariant gauges, unphysical degrees of freedom (longitudinal and time-like polarisations) also propagate. The effect of these unwanted degrees of freedom is cancelled by the ghost fields, which are coloured complex scalars obeying Fermi statistics. Unphysical degrees of freedom and the ghost fields can be avoided by choosing *axial gauges* (also called *physical gauges*). The axial gauge is defined by introducing an arbitrary vector n^μ with $p\cdot n \neq 0$, to impose the constraint

$$\mathcal{L}_{\rm GF} = -\frac{1}{2\alpha}\,(n^\mu A_\mu)^2\;\;, \tag{41}$$

which leads to

$$\Delta_{\mu\nu}(p,n) = \frac{-i}{p^2+i\,\varepsilon}\left(g_{\mu\nu} - \frac{p_\mu n_\nu + n_\mu p_\nu}{p\cdot n} + \frac{n^2\,p_\mu p_\nu}{(p\cdot n)^2}\right). \tag{42}$$

The so-called *light-cone gauge* is characterised by $n^2\,=\,0$. Note that the axial gauge propagator (42) satisfies

$$\Delta_{\mu\nu}\,(p,n)\;p^\mu = 0\;,\quad \Delta_{\mu\nu}\,(p,n)\;n^\mu = 0\;.$$

Thus, only 2 degrees of freedom propagate (transverse ones in the $n^\mu + p^\mu$ rest frame). The price to pay by choosing an axial gauge instead of a covariant one is that the propagator looks more complicated and that it diverges when p^μ becomes parallel to n^μ. In the light-cone gauge we have

$$\begin{aligned}
\Delta_{\mu\nu}(p,n) &= \frac{i}{p^2 + i\,\varepsilon}\, d_{\mu\nu}(p,n)\ , \\
d_{\mu\nu}(p,n) &= -g_{\mu\nu} + \frac{p_\mu n_\nu + n_\mu p_\nu}{p\cdot n} = \sum_{\lambda=1,2} \epsilon^\lambda_\mu(p) \left(\epsilon^\lambda_\nu(p)\right)^*\ ,
\end{aligned} \tag{43}$$

where $\epsilon^\lambda_\mu(p)$ is the polarisation vector of the gluon field with momentum p and polarisation λ. This means that only the two physical polarisations ($\lambda = 1, 2$) propagate. In contrast, in Feynman gauge, we have

$$d_{\mu\nu}(p) = -g_{\mu\nu} = \sum_{\lambda=0}^{3} \epsilon^\lambda_\mu(p) \left(\epsilon^\lambda_\nu(p)\right)^*\ , \tag{44}$$

where the polarisation sum also runs over non-transverse gluon polarisations. Unphysical polarisations that occur in loops will be cancelled by the corresponding loops involving ghost fields.

Faddeev–Popov ghost fields

The introduction of a gauge fixing constraint is achieved by inserting unity in the form

$$1 = \int \mathcal{D}\theta(x)\, \delta(G^a(A^\theta) - h^a(x))\ \det\left(\frac{\delta G^a(A^\theta)}{\delta\theta}\right) \tag{45}$$

into the generating functional $Z[J]$. A^θ denotes all fields which are equivalent through a gauge transformation involving the group parameter θ. In covariant gauges $h^a(x) = \partial^\mu A^a_\mu(x)$. The determinant $\det\left(\frac{\delta G^a(A^\theta)}{\delta\theta}\right) =: \Delta_{FP}(A)$ can be written as a functional integral over anti-commuting fields $\eta^a(x), \bar\eta^b(x)$,

$$\Delta_{FP}(A) = \int \mathcal{D}\bar\eta \mathcal{D}\eta\, e^{i\int d^4x d^4y\, \bar\eta^a(x) M_{ab}(x,y)\, \eta^b(y)} \quad \text{with} \quad M_{ab}(x,y) = \frac{\delta G^a(A^\theta(x))}{\delta\theta^b(y)}\ . \tag{46}$$

The fields $\bar\eta^a(x), \eta^b(x)$ are the so-called *Faddeev–Popov*-fields or *ghost* fields, they are complex scalar fields, which however obey Fermi-statistics, so they anti-commute. They cannot occur as external states because they do not have physical polarisations.

The additional term in the Lagrangian as a result of the procedure sketched above reads

$$\mathcal{L}_{FP} = \bar\eta_a\, M^{ab}\, \eta_b\ . \tag{47}$$

In covariant gauges, the operator M^{ab} (also called Faddeev–Popov matrix) is given by

$$M^{ab}_{\text{cov}} = \delta^{ab}\, \partial_\mu\partial^\mu + g_s\, f^{abc} A^c_\mu \partial^\mu\ . \tag{48}$$

Here we can see that in QED (or another Abelian gauge theory) the second term is absent, such that the Faddeev–Popov matrix does not depend on any field and therefore Δ_{FP} can be absorbed into the normalisation of the path integral, such that no ghost fields are needed in Abelian gauge theories.

In axial gauges, the Faddeev–Popov matrix becomes

$$M^{ab}_{\text{axial}} = \delta^{ab}\, n_\mu \partial^\mu + g_s\, f^{abc}\, n_\mu A^\mu_c \ , \tag{49}$$

such that, due to the gauge fixing condition $n \cdot A = 0$, the matrix is again independent of the gauge field and therefore can be absorbed into the normalisation, such that no ghost fields propagate.

Collecting all contributions, we finally have derived the full QCD Lagrangian

$$\boxed{\mathcal{L}_{QCD} = \mathcal{L}_{\text{YM}} + \mathcal{L}_q + \mathcal{L}_{GF} + \mathcal{L}_{FP} \ .} \tag{50}$$

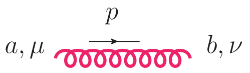
2.2.4 QCD Feynman rules

Feynman rules are something like a Lego brick box containing pieces that can be assembled to an expression describing a process such as the scattering of elementary particles or the decay of a particle. They can be derived from the interaction terms in the action, resp. the Lagrangian. There are also automated tools that can derive Feynman rules from a given Lagrangian, see e.g. [1, 26].

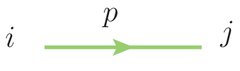
We will not derive the QCD Feynman rules from scratch, but just state them below.

Propagators: ($i\varepsilon$ prescription understood)

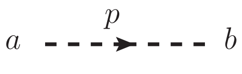
gluon propagator: $\Delta^{ab}_{\mu\nu}(p) = \delta^{ab}\, \Delta_{\mu\nu}(p)$



quark propagator: $\Delta^{ij}_q\,(p) = \delta^{ij}\, i\, \frac{\not{p}+m}{p^2-m^2}$

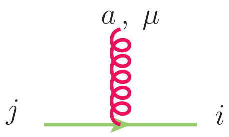


ghost propagator: $\Delta^{ab}\,(p) = \delta^{ab}\, \frac{i}{p^2}$

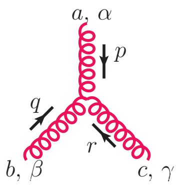


Vertices:

quark–gluon: $\Gamma^{\mu,\,a}_{gq\bar{q}} = -i\, g_s\, (t^a)_{ij} \gamma^\mu$

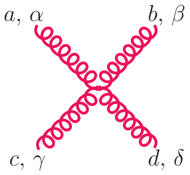


three-gluon: $\Gamma^{abc}_{\alpha\beta\gamma}(p,q,r) = -i\, g_s\, (F^a)_{bc}\, V_{\alpha\beta\gamma}(p,q,r)$

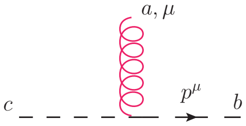


$$V_{\alpha\beta\gamma}(p,q,r) = (p-q)_\gamma g_{\alpha\beta} + (q-r)_\alpha g_{\beta\gamma} + (r-p)_\beta g_{\alpha\gamma}\,, \quad p^\alpha + q^\alpha + r^\alpha = 0$$

$$\text{four-gluon: } \Gamma^{abcd}_{\alpha\beta\gamma\delta} = -i\, g_s^2 \begin{bmatrix} +f^{xac}\, f^{xbd}\, (g_{\alpha\beta} g_{\gamma\delta} - g_{\alpha\delta} g_{\beta\gamma}) \\ +f^{xad}\, f^{xcb}\, (g_{\alpha\gamma} g_{\beta\delta} - g_{\alpha\beta} g_{\gamma\delta}) \\ +f^{xab}\, f^{xdc}\, (g_{\alpha\delta} g_{\beta\gamma} - g_{\alpha\gamma} g_{\beta\delta}) \end{bmatrix}$$



$$\text{ghost–gluon: } \Gamma^{\mu,\,a}_{g\eta\bar\eta} = -i\, g_s\, (F^a)_{bc}\, p^\mu$$



External lines are represented by the spinors for incoming and outgoing fermions with momentum p and spin s and polarisation vectors for vector bosons with polarisation λ (see Fig. 14).

- outgoing fermion: $\bar{u}(p,s)$
- incoming fermion: $u(p,s)$
- outgoing vector boson: $\epsilon^*_\mu(p,\lambda)$
- outgoing antifermion: $v(p,s)$
- incoming antifermion: $\bar{v}(p,s)$
- incoming vector boson: $\epsilon_\mu(p,\lambda)$.

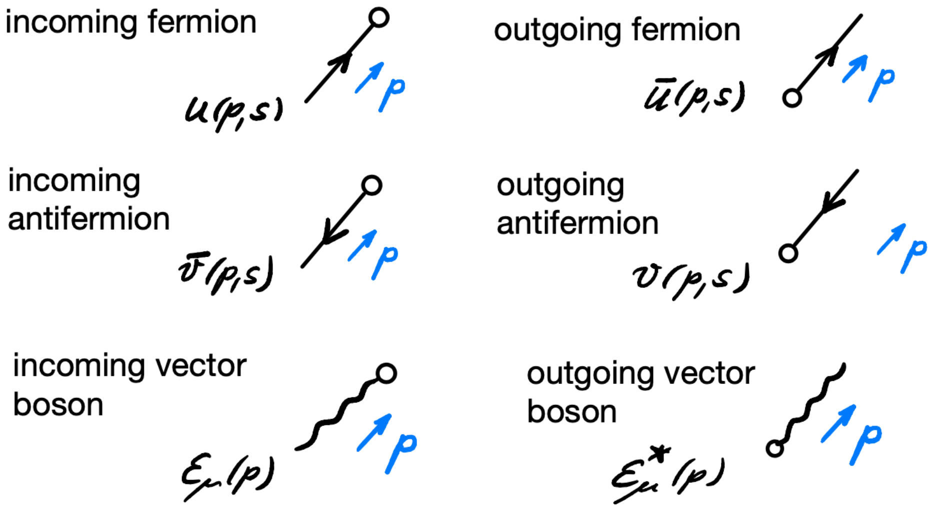


Fig. 14: Conventions for spinors with momentum p and spin s and polarisation vectors $\epsilon_\mu(p,\lambda)$.

3 Amplitudes and cross sections

3.1 Tree-level amplitudes

3.1.1 Partonic cross sections

The partonic cross section $\hat{\sigma}_{ij}$ in eq. (2) can be calculated order by order in perturbation theory. It contains the modulus of the scattering or decay matrix element, $|\mathcal{M}|^2$, which encodes the fundamental interactions derived from the Lagrangian.

For a reaction $p_i + p_j \to p_1 + \ldots + p_n$, the reaction rate can be calculated according to Fermi's golden rule based on the transition matrix element $|\mathcal{M}|^2$. We have

$$\begin{aligned} \mathrm{d}\hat{\sigma} &= \frac{J}{\text{flux}} \cdot |\mathcal{M}|^2 \cdot \mathrm{d}\Phi_n \;, \\ \text{flux} &= 4\sqrt{(p_i \cdot p_j)^2 - m_i^2 m_j^2} \;. \end{aligned} \tag{51}$$

Assuming massless incoming particles and calculating in the centre-of-mass frame of $p_i + p_j$, we find flux $= 4p_i \cdot p_j = 2\hat{s}$, with $\hat{s} = (p_i + p_j)^2$. The quantity $J = 1/j!$ is a statistical factor to be included for each group of j identical particles in the final state. The phase-space volume spanned by the final-state particles is denoted by $\mathrm{d}\Phi_n$, it will be considered in more detail in section 4.1.

For a decay process $Q \to p_1 + \ldots + p_n$ we have

$$\mathrm{d}\Gamma = \frac{J}{2\sqrt{Q^2}} \cdot |\mathcal{M}|^2 \cdot \mathrm{d}\Phi_n \;. \tag{52}$$

If the polarisations or spins of the final state particles are not measured, we sum over all possible polarisations/spins in the final state. Colour in the final state cannot be measured, so we also have to sum over all colours in the final state. Furthermore, we average over all possible colours, polarisations and spins in the initial state. The matrix element is then given by

$$|\mathcal{M}|^2 \to \overline{\sum}|\mathcal{M}|^2 = \left(\prod_{\text{initial}} \frac{1}{N_{\text{pol}} N_{\text{col}}}\right) \sum_{\text{pol,col}} |\mathcal{M}|^2 \;, \tag{53}$$

where for quarks $N_{\text{col}} = N_c$, for gluons $N_{\text{col}} = N_c^2 - 1$, and $N_{\text{pol}} = 2$ for both, quarks and gluons, as long as we stay in four space–time dimensions. The expression $\overline{\sum}|\mathcal{M}|^2$ is often just written as $|\overline{\mathcal{M}}|^2$.

3.1.2 Total hadronic cross section

For hadron colliders, the cross section σ at a centre-of-mass energy $\sqrt{s}$ can be expressed in terms of the partonic cross section and a luminosity function

$$\mathcal{L}_{ij}(x_1, x_2, \mu_f) = f_{i/p_a}(x_1, \mu_f) f_{j/p_b}(x_2, \mu_f) \;, \tag{54}$$

where $f_i(x, \mu_f)$ is the parton distribution function of a parton with momentum fraction x and flavour i (including gluons) and μ_f is the factorisation scale, such that

$$\begin{aligned} \sigma(s) &= \int_{\hat{s}_{min}}^{s} \mathrm{d}\hat{s} \int_0^1 \mathrm{d}x_1 \int_0^1 \mathrm{d}x_2\, \delta(\hat{s} - x_1 x_2 s) \sum_{i,j} \mathcal{L}_{ij}(x_1, x_2, \mu_f) \int \mathrm{d}\hat{\sigma}_{ij}(x_1, x_2, \mu_r, \mu_f) \\ &= \int_{\tau_{min}}^{1} \mathrm{d}\tau \int_\tau^1 \frac{\mathrm{d}x}{x} \sum_{ij} f_i(x, \mu_f)\, f_j(\frac{\tau}{x}, \mu_f)\, \sigma_{ij}(\hat{s} = \tau s) \end{aligned} \tag{55}$$

In the second line the replacement $x_{1,2} = \sqrt{\tau} e^{\pm y}$, $\mathrm{d}x_1 \mathrm{d}x_2 = \mathrm{d}\tau \mathrm{d}y$ has been made.

3.2 Running couplings and masses

In this section we would like to explain how it arises that theoretical predictions depend in general on at least one unphysical scale, the so-called *renormalisation scale* μ. In the case of hadronic initial-state particles, there is also a *factorisation scale* μ_f involved. There can be even more unphysical scales, like fragmentation scales in the modelling of the fragmentation of final-state particles into hadrons, parton shower matching scales, or resummation scales.

Let us first motivate how the dependence on a renormalisation scale arises. We mentioned already that the strong coupling, defined as $\alpha_s = g_s^2/(4\pi)$, is not really a constant. To leading order in the perturbative expansion, it obeys the relation

$$\alpha_s(Q^2) = \frac{\alpha_s(\mu^2)}{1 + b_0\, t\, \alpha_s(\mu^2)} \quad , \quad ; t = \ln \frac{Q^2}{\mu^2} \; . \tag{56}$$

where Q^2 is a generic energy scale, for example the centre-of-mass energy of a scattering process. The coefficient b_0 is given by

$$b_0 = \frac{1}{4\pi}\left(\frac{11}{3}\, C_A - \frac{4}{3} T_R\, N_F\right) \; , \tag{57}$$

where N_F is the number of flavours in the quark loops contributing to the gluon self-energy. Note that $b_0 > 0$ for $N_F < 11/2\, C_A$.

Where does the running of the coupling come from? It is closely linked to renormalisation, which introduces the *renormalisation scale* μ. Before we enter into the technicalities, let us look at a physical observable, for example the R-ratio which we encountered already,

$$R(s) = \frac{\sigma(e^+e^- \to \text{ hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)} \; , \tag{58}$$

assuming that the energy exchanged in the scattering process is much larger than Λ_{QCD}.

To get a precise result, we should include quantum corrections, for example diagrams where virtual gluons are exchanged, such as the ones shown in Fig. 15. The perturbative expansion for R can be written as

$$\begin{aligned} R(s) &= K_{QCD}(s)\, R_0 \; , \quad R_0 = N_c \sum_f Q_f^2\, \theta(s - 4m_f^2) \; , \\ K_{QCD}(s) &= 1 + \frac{\alpha_s(\mu^2)}{\pi} + \sum_{n\geq 2} C_n \left(\frac{s}{\mu^2}\right) \left(\frac{\alpha_s(\mu^2)}{\pi}\right)^n \; . \end{aligned} \tag{59}$$

The coefficients C_n can be calculated perturbatively, see e.g. Refs. [7,52,61] for results of order α_s^4 (five loops).

However, when trying to perform the loop integrals over a loop momentum k, we will find that they diverge for $k \to \infty$ in 4 space–time dimensions (see Section 3.4.1), therefore we need to introduce a regulator to be able to calculate them at all; usually dimensional regularisation is employed, calculating the integrals in $D = 4 - 2\epsilon$ space–time dimensions. For the argument to make here we instead use an arbitrary cutoff scale Λ_{UV} for the upper integration boundary. If we carried through the calculation, we

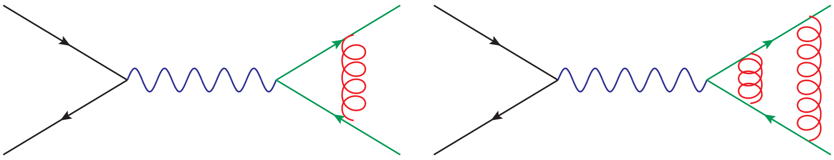


Fig. 15: Examples of one- and two-loop diagrams contributing to $e^+e^- \to q\bar{q}$.

would see that the dependence on the cutoff in the left diagram of Fig. 15 cancels, which is a consequence of the Ward Identity. However, if we go one order higher in α_s, calculating diagrams like the one on the right-hand side in Fig. 15, the cutoff-dependence does not cancel anymore. We obtain

$$K_{QCD}(s) = 1 + \frac{\alpha_s}{\pi} + \left(\frac{\alpha_s}{\pi}\right)^2 \left[c + b_0\pi \, \log\left(\frac{\Lambda_{UV}^2}{s}\right)\right] + \mathcal{O}(\alpha_s^3) \;, \tag{60}$$

where c is a constant. So our result is infinite as we take the limit $\Lambda_{UV} \to \infty$. However, we did not claim that α_s is the coupling we measure. In fact, it is the "bare" coupling, also denoted as α_s^0, which appears in Eq. (60), and we can absorb the infinity in the bare coupling to arrive at the renormalised coupling, which is identified with the one we measure. In our case, this looks as follows. We define

$$\alpha_s(\mu) = \alpha_s^0 + b_0 \log\left(\frac{\Lambda_{UV}^2}{\mu^2}\right) \alpha_s^2 \;, \tag{61}$$

then replace α_s^0 by $\alpha_s(\mu)$ and drop consistently all terms of order α_s^3. In other words, the divergence arising in the bare coupling α_s^0 when calculating quantum corrections is cancelled by the second term, the *counter term*. This leads to

$$K_{QCD}^{\text{ren}}(\alpha_s(\mu), \mu^2/s) = 1 + \frac{\alpha_s(\mu)}{\pi} + \left(\frac{\alpha_s(\mu)}{\pi}\right)^2 \left[c + b_0\pi \, \log\left(\frac{\mu^2}{s}\right)\right] + \mathcal{O}(\alpha_s^3) \;. \tag{62}$$

K_{QCD}^{ren} is finite, but now it depends on the scale μ, both explicitly and implicitly through $\alpha_s(\mu)$. However, the hadronic R-ratio is a physical quantity and therefore cannot depend on the arbitrary scale μ. The dependence of K_{QCD} on μ is an artefact of the truncation of the perturbative series after the order α_s^2.

Renormalisation group and asymptotic freedom

Since the hadronic R-ratio $R^{\text{ren}} = R_0\, K_{QCD}^{\text{ren}}$ cannot depend on μ (if it was calculated to all orders in perturbation theory), its total derivative w.r.t. μ must vanish:

$$\mu^2 \frac{\mathrm{d}}{\mathrm{d}\mu^2} R^{\text{ren}}(\alpha_s(\mu), \mu^2/Q^2) = \; 0 \; = \left(\mu^2 \frac{\partial}{\partial \mu^2} + \mu^2 \frac{\partial \alpha_s}{\partial \mu^2} \frac{\partial}{\partial \alpha_s}\right) R^{\text{ren}}(\alpha_s(\mu), \mu^2/Q^2) \;. \tag{63}$$

Equation (63) is an example of a so-called *renormalisation group equation (RGE)*. Introducing the abreviations

$$t = \ln \frac{Q^2}{\mu^2} \;, \qquad \beta(\alpha_s) = \mu^2 \frac{\partial \alpha_s}{\partial \mu^2} \;, \tag{64}$$

the RGE becomes

$$\left(-\frac{\partial}{\partial t} + \beta(\alpha_s) \frac{\partial}{\partial \alpha_s}\right) R^{\text{ren}} = 0 \;. \tag{65}$$

This first-order partial differential equation can be solved by implicitly defining a function $\alpha_s(Q^2)$, the *running coupling*, such that

$$t = \int_{\alpha_s}^{\alpha_s(Q^2)} \frac{\mathrm{d}x}{\beta(x)} \,, \quad \text{with} \quad \alpha_s \equiv \alpha_s(\mu^2) \,, \tag{66}$$

where μ^2 above is fixed to an arbitrary value. We can solve Eq. (66) perturbatively using an expansion of the β-function

$$\beta(\alpha_s) = -b_0\alpha_s^2 \left[1 + \sum_{n=1}^{\infty} b_n\, \alpha_s^n\right] , \tag{67}$$

where $b_0 = \frac{\beta_0}{4\pi}$ and $b_0 b_1 = \frac{\beta_1}{(4\pi)^2}$, etc. Explicitly, up to NNLO:

$$\mu^2 \frac{\mathrm{d}\alpha_s(\mu)}{\mathrm{d}\mu^2} = -\alpha_s(\mu) \left[\beta_0 \left(\frac{\alpha_s(\mu)}{2\pi}\right) + \beta_1 \left(\frac{\alpha_s(\mu)}{2\pi}\right)^2 + \beta_2 \left(\frac{\alpha_s(\mu)}{2\pi}\right)^3 + \mathcal{O}(\alpha_s^4)\right] .$$

The four-loop coefficients are known since some time [76], the β-function at five loops has been calculated only recently [7,22,28,52,61]. The first three coefficients (in the $\overline{\mathrm{MS}}$-scheme) are

$$\begin{aligned}
\beta_0 &= \frac{11\,C_A - 4T_R N_F}{6} \,, \\
\beta_1 &= \frac{17\,C_A^2 - 10 C_A T_R N_F - 6 C_F T_R N_F}{6} \,, \\
\beta_2 &= \frac{1}{432}\big(2857 C_A^3 + 108 C_F^2 T_R N_F - 1230 C_F C_A T_R N_F - 2830 C_A^2 T_R N_F \\
&\quad + 264 C_F T_R^2 N_F^2 + 316 C_A T_R^2 N_F^2\big) \,.
\end{aligned} \tag{68}$$

Truncating the series Eq. (67) at leading order leads to the simple solution given already in Eq. (56),

$$\begin{aligned}
Q^2 \frac{\partial \alpha_s}{\partial Q^2} = \frac{\partial \alpha_s}{\partial\, t} = -b_0 \alpha_s^2 \;\Rightarrow\; -\frac{1}{\alpha_s(Q^2)} + \frac{1}{\alpha_s(\mu^2)} = -b_0\, t \\
\Rightarrow\; \alpha_s(Q^2) = \frac{\alpha_s(\mu^2)}{1 + b_0\, t\, \alpha_s(\mu^2)} \,.
\end{aligned} \tag{69}$$

Eq. (69) implies that

$$\alpha_s(Q^2) \overset{Q^2 \gg \mu^2}{\longrightarrow} \frac{1}{b_0 t} \overset{Q^2 \to \infty}{\longrightarrow} 0 \,. \tag{70}$$

This behaviour is called *asymptotic freedom*: the larger Q^2, the smaller the coupling, so at very high energies (small distances), the quarks and gluons can be treated as if they were free particles. The behaviour of α_s as a function of Q^2 is illustrated in Fig. 16 including measurements at different energies. Note that the sign of b_0 is positive for QCD (see eq. (57)), while its analogue in QED is negative. The decrease of the coupling $\alpha_s(Q^2)$ with increasing Q^2 can be related to “screening and anti-screening” effects of colour charges, to be contrasted to the screening of electric charges due to vacuum polarisation in the QED case. The gluon loops, leading to the C_A-term in b_0, have the opposite sign relative to the quark loops, and the effect of the gluon loops is dominating: the positive term in b_0 is larger than the negative term. It can be proven that, in four space–time dimensions, only non-Abelian gauge theories

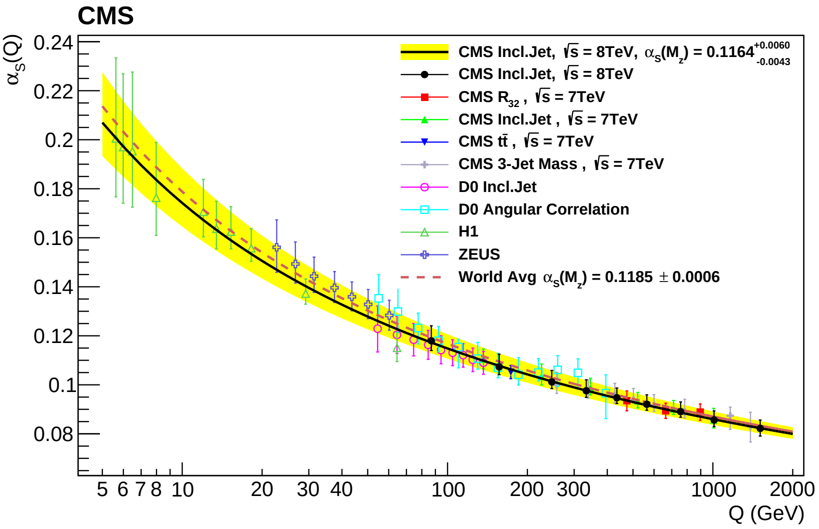


Fig. 16: The running coupling $\alpha_s(Q^2)$. Figure from Ref. [54].

can be asymptotically free. For the discovery of asymptotic freedom in QCD [47, 69], Gross, Politzer, and Wilczek got the Nobel Prize (much later, in 2004).

The QCD Λ-parameter

Instead of the arbitrary scale μ in Eq. (66) one can also use a characteristic scale, the scale where the coupling would diverge or at least becomes so strong that perturbation theory makes no sense anymore. This scale is the QCD Λ-parameter. It is defined by

$$\log\left(\frac{Q^2}{\Lambda^2_{QCD}}\right) = -\int_{\alpha_s(Q^2)}^{\infty} \frac{\mathrm{d}x}{\beta(x)} = \int_{\alpha_s(Q^2)}^{\infty} \frac{\mathrm{d}x}{b_0 x^2(1 + b_1 x + \ldots)} \tag{71}$$

Keeping only the leading term we find the familiar form

$$\alpha_s(Q^2) = \frac{1}{b_0 \log\left(Q^2/\Lambda^2_{QCD}\right)} \,, \tag{72}$$

Using $L = \log(Q^2/\Lambda^2_{QCD})$ yields the following solution up to NNLO,

$$\alpha_s(Q^2) = \frac{4\pi}{\beta_0 L}\left(1 - \frac{\beta_1}{\beta_0^2}\frac{\log L}{L} + \frac{1}{\beta_0^2 L^2}\left(\frac{\beta_1^2}{\beta_0^2}\left(\log^2 L - \log L - 1\right) + \frac{\beta_2}{\beta_0}\right)\right) \,. \tag{73}$$

Note that Λ_{QCD} depends on the number of flavours because $N_F = 4$ for $m_c < Q < m_b$, $N_F = 5$ for $m_b < Q < m_t$, etc.

Running quark masses

In the derivation of the RGE in Eq. (63) above, we have assumed that the observable R does not depend on other couplings or parameters, such as quark masses. However, if the considered observable depends on masses, the latter also require renormalisation. Introducing the renormalisation constant Z_m, we have

$$\begin{aligned} \mu\frac{d\,m^{(0)}}{d\mu} &= 0\ ,\ \ m^{(0)} = Z_m\, m \\ \Rightarrow \mu\,\frac{d\,Z_m}{d\mu}\cdot m + \mu\cdot Z_m\frac{dm}{d\mu} &= 0\ . \end{aligned} \tag{74}$$

The *mass anomalous dimension* γ_m is defined by

$$\mu^2\frac{\partial m}{\partial\mu^2} = -\gamma_m(\alpha_s(\mu^2))\, m(\mu^2)\ , \tag{75}$$

where the minus sign pulled out in front of γ_m is convention.

This naturally leads to renormalisation group equations that are extended to include mass renormalisation, which will lead to running quark masses:

$$\left(\mu^2\frac{\partial}{\partial\mu^2} + \beta(\alpha_s)\frac{\partial}{\partial\alpha_s} - \gamma_m(\alpha_s)m\frac{\partial}{\partial\, m}\right) R\left(\frac{Q^2}{\mu^2}, \alpha_s, \frac{m}{Q}\right) = 0\ . \tag{76}$$

In a perturbative expansion we can write the mass anomalous dimension as

$$\gamma_m(\alpha_s)\ =\ c_0\,\alpha_s\left(1 + \sum_n c_n\alpha_s^n\right)\ .$$

The coefficients are known up to c_4, i.e. $\mathcal{O}(\alpha_s^5)$ [5, 6, 49, 60].

Even though the top quark mass is not an observable, a more commonly used renormalisation scheme for top quark masses is the on-shell renormalisation scheme, where the idea is to preserve the on-shell mass of a particle, which corresponds to the zero of the real part of the inverse propagator. However, the on-shell mass of the top quark suffers from so-called renormalon contributions, which affect the convergence properties of the perturbative expansion. The conversion between the pole mass M_t and the $\overline{\rm MS}$ mass $m_t(\mu_m)$ is given by [18]

$$M_t = m_t(\mu_m)\, d(m_t(\mu_m), \mu_m) = m_t(\mu_m)\left(1 + \sum_{k=1}^{\infty}\left(\frac{\alpha_s(\mu_m)}{\pi}\right)^k d^{(k)}(\mu_m)\right). \tag{77}$$

The first two perturbative coefficients $d^{(1)}$ and $d^{(2)}$ in Eq. (77) have the values [37, 44]

$$\begin{aligned} d^{(1)}(\mu_m) &= \frac{4}{3} + L_{\mu_m}\ , \\ d^{(2)}(\mu_m) &= \frac{307}{32} + 2\,\zeta_2 + \frac{2}{3}\,\zeta_2\,\ln 2 - \frac{1}{6}\,\zeta_3 + \frac{509}{72}\,L_{\mu_m} + \frac{47}{24}\,L^2_{\mu_m} \\ &\quad - \left(\frac{71}{144} + \frac{1}{3}\,\zeta_2 + \frac{13}{36}\,L_{\mu_m} + \frac{1}{12}\,L^2_{\mu_m}\right) N_F\ , \end{aligned} \tag{78}$$

where

$$L_{\mu_m} = 2\ln(\mu_m/m_t(\mu_m))\ . \tag{79}$$

Higher-order coefficients were computed in Refs. [23, 36, 63, 66].

3.3 Scale uncertainties

Let us consider an observable O, calculated in perturbation theory to order α_s^{N+k},

$$O = \sum_{n=0}^{N} C_n(\mu_r)\alpha_s^{n+k}(\mu_r) \;,$$

where k is the power of α_s of the leading order cross section. From the perturbative solution of the RGE we can derive how the physical quantity $O^{(N)}(\mu)$, truncated at order N in perturbation theory, changes with the renormalisation scale μ:

$$\frac{\mathrm{d}}{\mathrm{d}\log(\mu^2)}\, O^{(N)}(\alpha_s(\mu)) = \beta(\alpha_s)\frac{\partial O^{(N)}}{\partial \alpha_s} \sim \mathcal{O}\left(\alpha_s^{N+1}(\mu)\right) \;, \tag{80}$$

where the order $\alpha_s^{N+1}(\mu)$ arises because $\beta(\alpha_s) = -b_0\alpha_s^2 + \mathcal{O}(\alpha_s^3)$. This means that, the more higher-order coefficients C_n we can calculate, the weaker the dependence of the result on the unphysical scale μ^2 will be. Therefore, the dependence on the scale is used to estimate the uncertainty of a result stemming from missing higher orders. Usually the scale is varied by a factor of two up and down.

In hadronic collisions there is another scale, the factorisation scale μ_f, which comes from the factorisation of initial-state infrared singularities. It also needs to be taken into account when assessing the uncertainty of a theoretical prediction. Varying both μ_r and μ_f simultaneously in the same direction can lead to accidental cancellations and hence an underestimation of the perturbative uncertainties. Therefore, in the presence of both μ_r and μ_f, usually so-called *7-point scale variations* are performed, which means $\mu_{r,f} = c_{r,f}\mu_0$, where $c_r, c_f \in \{2, 1, 0.5\}$ and where the extreme variations $(c_r, c_f) = (2, 0.5)$ and $(c_r, c_f) = (0.5, 2)$ have been omitted.

Still, the question remains what to choose for the central scale μ_0. A convenient choice is a scale where the higher-order corrections are small, i.e. a scale showing good "perturbative stability".

Let us now look at a few examples where such scale variations do not capture the true uncertainties. First some prelimiary remarks. If there is only one scale μ involved, the scale dependence of an observable is given through $\alpha_s(\mu)$, and we can use the renormalisation group equation to move from the result at a scale μ_0 to a result at a different scale. For the observable O, known to order α_s^{N+k}, we can use the requirement $\mathrm{d}O/\mathrm{d}\ln\left(\frac{\mu_r^2}{\mu_0^2}\right) = 0$ and Eq. (67) to derive how O changes with a change of scale, leading to

$$O = \alpha_s^k(\mu_r)\left\{C_0 + \left(C_1 + b_0 C_0 \ln\left(\frac{\mu_r^2}{\mu_0^2}\right)\right)\alpha_s(\mu_r) + \mathcal{O}(\alpha_s^2)\right\} \;. \tag{81}$$

Variations of μ_r will change the C_0-part of the $\mathcal{O}(\alpha_s^{k+1})$ term, however the magnitude of C_1 can only be determined by direct calculation.

For some processes, C_1 (and C_2) turned out to be pretty large, and the scale uncertainty bands obtained from 7-point scale variations do not (fully) overlap between the different orders. One such example is Higgs production in gluon fusion, known to order N^{3}LO. Figure 17 (left) shows a slow stabilisation of the scale dependence with increasing perturbative order, and the higher-order corrections are very large. The standard scale uncertainty bands are shown in Fig. 17 (right). It is obvious that the LO scale variation band would be a very poor measure of the uncertainty due to missing higher orders. Among the reasons for the large K-factors (i.e. the relative size of the higher-order corrections), in particular the NLO K-factor, are large colour factors and new partonic channels opening up. Furthermore, the behaviour of the scale uncertainty bands can depend sensitively on the definition of the central scale.

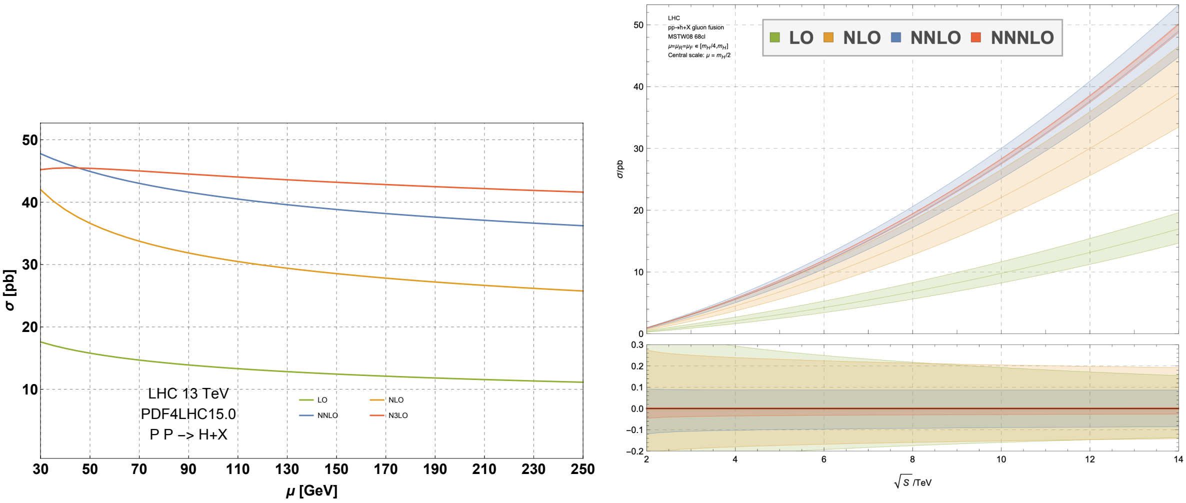


Fig. 17: Left: Higgs production in gluon fusion, stabilisation of the scale dependence. Figure from Ref. [67]. Right: Scale uncertainty bands for Higgs production in gluon fusion. Figure from Ref. [4].

Usually one can see that the perturbative series stabilises at latest between NNLO and N^{3}LO. However, there are exceptions as well, an example is charged-current Drell–Yan production, calculated to N^{3}LO in Refs. [15, 21, 34]. With a central scale of $Q = 100\,\text{GeV}$, NNLO PDFs and 7-point scale variations, the NNLO and N^{3}LO uncertainty bands do not overlap, see Fig. 18. Looking at the variation of the factorisation scale separately, the NNLO μ_f-uncertainty band is found to be accidentally small; this has been traced back to cancellations between different partonic channels.

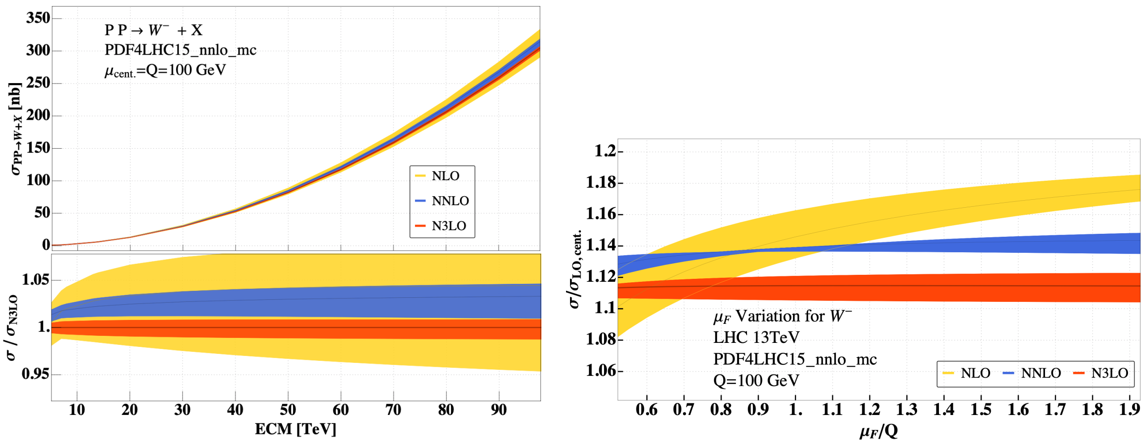


Fig. 18: Charged-current Drell–Yan production, left: cross section with scale bands for combined renormalisation and factorisation scale dependence. Right: factorisation scale dependence shown separately. Figure from Ref. [34].

3.4 Basics of NLO QCD calculations

3.4.1 Dimensional regularisation

Tree-level results in QCD are usually not accurate enough to match the current experimental precision and suffer from large scale uncertainties. When calculating higher orders, we encounter non-integrable singularities: ultraviolet (UV) singularities and infrared (IR) singularities, the latter are due to soft or collinear massless particles. Therefore the introduction of a *regulator* is necessary.

Let us first have a look at UV singularities: The expression for the one-loop two-point function shown in Fig. 19 naively would be

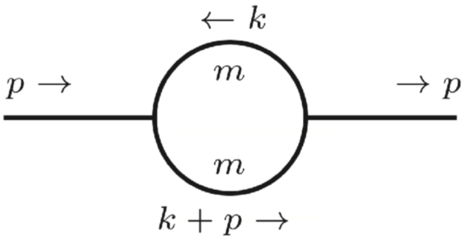


Fig. 19: One-loop two-point function I_2 ("bubble").

$$I_2 = \int_{-\infty}^{\infty} \frac{d^4k}{(2\pi)^4} \frac{1}{[k^2 - m^2 + i\delta][(k+p)^2 - m^2 + i\delta]} \,. \tag{82}$$

If we are only interested in the behaviour of the integral for $|k| \to \infty$ we can neglect the masses, transform to polar coordinates and obtain

$$I_2 \sim \int \mathrm{d}\Omega_3 \int_0^{\infty} \mathrm{d}|k| \frac{|k|^3}{|k|^4} \,. \tag{83}$$

This integral is clearly not well-defined. If we introduce an upper cutoff Λ (and a lower limit $|k|_{\min}$ because we neglected the masses and p^2, which would serve as an IR regulator), the integral is regulated:

$$I_2 \sim \int_{|k|_{\min}}^{\Lambda} \mathrm{d}|k| \frac{1}{|k|} \sim \log\left(\frac{\Lambda}{|k|_{\min}}\right) \,. \tag{84}$$

The integral has a logarithmic UV divergence for $\Lambda \to \infty$. The regularisation with a cut-off Λ is problematic since it is neither a Lorentz invariant nor a gauge invariant way to regulate integrals over loop momenta. A regularisation method which preserves these symmetries is *dimensional regularisation*.

Dimensional regularisation has been introduced in 1972 by ‘t Hooft and Veltman [74], and by Bollini and Giambiagi [12], as a method to regularise UV divergences in a gauge invariant way, thus completing the proof of renormalisability. The idea is to work in $D = 4 - 2\epsilon$ space–time dimensions. Divergences for $D \to 4$ will appear as poles in $1/\epsilon$. This means that objects such as momenta, polarisation vectors and the metric tensor live in a D-dimensional space. The γ-algebra also has to be extended to D dimensions, however, how to treat unphysical objects is not unique. There are several *regularisation schemes* within dimensional regularisation. For example, when doing a calculation in supersymmetry, it is inconvenient to use a scheme where massless bosons have $D - 2$ polarisation states while massless fermions have two polarisation states. Of course, all the different schemes must lead to the same result for physical quantities.

An important feature of dimensional regularisation is that, apart from UV singularities, it also regulates IR singularities, i.e. divergences occurring when massless particles become soft and/or collinear. Ultraviolet divergences occur for loop momenta $k \to \infty$, so in general the UV behaviour becomes better for $\epsilon > 0$, while the IR behaviour gets better for $\epsilon < 0$. Certainly we cannot have $D < 4$ and $D > 4$ at the same time. What is formally done is to first assume the IR divergences are regulated in some other way, e.g. by assuming all external legs are off-shell or by introducing a small mass for all massless particles. In this case all poles in $1/\epsilon$ will be of UV nature and renormalisation can be performed. After renormalisation we can analytically continue to the whole complex D-plane, in particular to $\text{Re}(D) > 4$. If we now remove the auxiliary IR regulator, the IR divergences will show up as $1/\epsilon$ poles (this is however not commonly done in practice, where all poles just show up as $1/\epsilon$ poles, and after UV renormalisation, the remaining poles must be of IR origin).

In dimensional regularisation, slight changes to the Feynman rules are to be made: we multiply the couplings in the Lagrangian by a factor μ^ϵ: $g \to g\mu^\epsilon$, where μ is an arbitrary mass scale. This ensures that each term in the Lagrangian has the correct mass dimension while retaining a dimensionless coupling in which we perform the perturbative expansion. The momentum integration involves $\int \frac{d^D k}{(2\pi)^D}$ for each loop.

We will not dive more deeply into the subject of loop calculations here, but rather discuss some general features of NLO calculations below.

4 Soft and collinear emissions, Jets

4.1 Cancellation of infrared singularities

Next-to-leading order calculations consist of several parts, which can be classified as virtual corrections (containing usually one loop), real corrections (radiation of extra particles relative to the leading order), and subtraction terms to deal with singularities. In the following we will assume that the virtual corrections already include UV renormalisation, such that the subtraction terms only concern the subtraction of the infrared (IR) singularities. IR singularities occur when a massless particle becomes soft (low energy) or when two massless particles become collinear to each other.

We will consider the next-to-leading order in an expansion in the strong coupling constant α_s. For electroweak corrections, the general structure is similar. The real and virtual contributions to the simple example $\gamma^* \to q\bar{q}$ (the hadronic part of $e^+e^- \to q\bar{q}$) are shown in Fig. 20.

If $\mathcal{M}_0$ is the leading-order amplitude and $\mathcal{M}_{\text{virt}}, \mathcal{M}_{\text{real}}$ are the virtual and real NLO amplitudes as shown in Fig. 20, the corresponding cross section is given by

$$\sigma^{NLO} = \underbrace{\int d\phi_2\, |\mathcal{M}_0|^2}_{\sigma^{LO}} + \int_R d\phi_3\, |\mathcal{M}_{\text{real}}|^2 + \int_V d\phi_2\, 2\text{Re}\left(\mathcal{M}_{\text{virt}}\mathcal{M}_0^*\right) \;. \tag{85}$$

The sum of the integrals $\int_R$ and $\int_V$ above is finite. However, this is not true for the individual contributions. The real part contains divergences due to soft and collinear radiation of massless particles. While

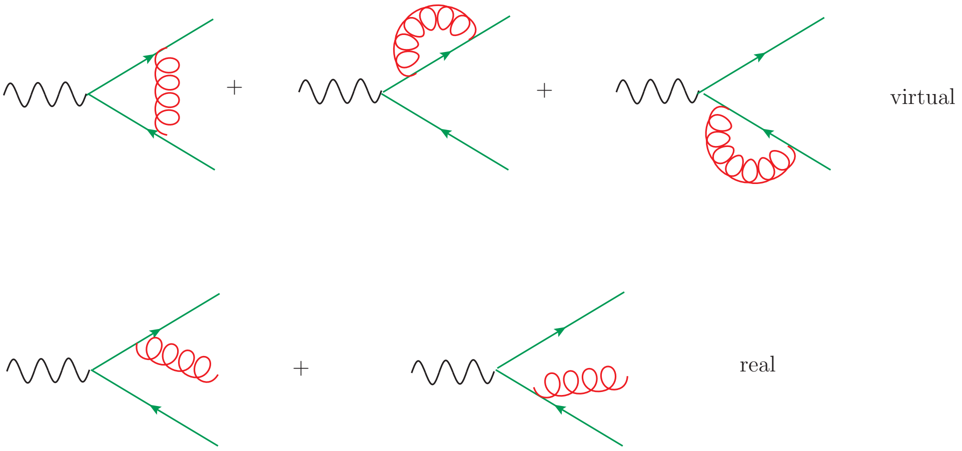


Fig. 20: The real and virtual NLO QCD contributions to $\gamma^* \to q\bar{q}$.

$\mathcal{M}_{\text{real}}$ itself is a tree level amplitude and thus finite, the divergences show up upon integration over the phase space $\mathrm{d}\Phi_3$. In $\int_V$, the phase space is the same as for the Born amplitude, but the loop integrals in $\mathcal{M}_{\text{virt}}$ contain explicit IR singularities.

Let us anticipate the answer, which we will (partly) calculate later. We find:

$$\begin{aligned}\sigma_R &= \sigma^{\text{Born}} \tilde{H}(\epsilon)\, C_F \frac{\alpha_s}{2\pi}\left(\frac{2}{\epsilon^2}+\frac{3}{\epsilon}+\frac{19}{2}\right)\,, \\ \sigma_V &= \sigma^{\text{Born}} H(\epsilon)\, C_F \frac{\alpha_s}{2\pi}\left(-\frac{2}{\epsilon^2}-\frac{3}{\epsilon}-8\right)\,,\end{aligned} \tag{86}$$

where $H(\epsilon) = \left(\frac{4\pi\mu^2}{-Q^2}\right)^{\epsilon} \frac{\Gamma(1+\epsilon)\Gamma^2(1-\epsilon)}{\Gamma(1-2\epsilon)}$ and $\tilde{H}(\epsilon) = H(\epsilon) + \mathcal{O}(\epsilon^3)$. The exact ϵ-dependence of $H(\epsilon) = 1 + \mathcal{O}(\epsilon)$ is irrelevant after summing up real and virtual contributions, because the poles in ϵ all cancel. This must be the case according to the **KLN theorem** (Kinoshita–Lee–Nauenberg) [55,58]. It says that

IR singularities must cancel when summing the transition rate over all degenerate (initial and final) states.

In our example, we do not have initial state singularities. However, in the final state we can have a massless quark accompanied by a soft gluon, or a collinear quark–gluon pair. Such a state cannot be distinguished from just a quark state, and therefore these two configurations are "degenerate". Only when summing over all the final-state multiplicities contributing to the cross section at a given order in α_s, the divergences cancel. Another way of stating this is by looking at the squared amplitude at order α_s and considering all cuts, see Fig. 21 (self-energy contributions, which are zero for massless quarks, are not shown). The KLN theorem states that *the sum of all diagrams resulting from cuts that lead to physical final states is free of IR poles.*

The cancellations between $\int_R$ and $\int_V$ in Eq. (85) are non-trivial, because the phase-space integrals contain a different number of particles in the final state and are thus of different dimensionality.

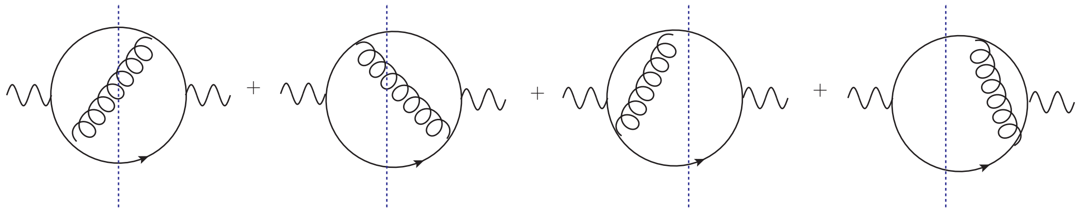


Fig. 21: The sum over cuts of the amplitude squared shown above is finite according to the KLN theorem.

Infrared safety

If we want to calculate a prediction for a certain observable, based on an n-particle final state, we need to multiply the amplitude by a *measurement function* $J(p_1 \dots p_n)$. The measurement function can contain for example a jet definition, or the definition of an event shape observable, or it defines observables such as the transverse momentum distribution of a final state particle. Schematically, the structure of the NLO cross section is the following. In the real radiation part, we have $n+1$ particles in the final state. Therefore the measurement function for the real radiation part depends on $n+1$ particles, while for the Born and virtual parts it only depends on n particles. Let us consider the case where we have an IR pole if the variable x, describing for example the energy of an extra gluon with momentum p_{n+1} in the real radiation part, goes to zero. If we define

$$\begin{aligned}
\mathcal{B}_n &= \int \mathrm{d}\phi_n \, |\mathcal{M}_0|^2 = \int \mathrm{d}\phi_n B_n \\
\mathcal{V}_n &= \int \mathrm{d}\phi_n \, 2\mathrm{Re}\,(\mathcal{M}_{\mathrm{virt}}\mathcal{M}_0^*) = \int \mathrm{d}\phi_n \frac{V_n}{\epsilon} \\
\mathcal{R}_n &= \int \mathrm{d}\phi_{n+1} \, |\mathcal{M}_{\mathrm{real}}|^2 = \int \mathrm{d}\phi_n \int_0^1 \mathrm{d}x \, x^{-1-\epsilon} \, R_n(x)
\end{aligned} \tag{87}$$

and a measurement function $J(p_1 \dots p_n, p_{n+1})$ we have

$$\sigma^{NLO} = \int \mathrm{d}\phi_n \left\{ \left(B_n + \frac{V_n}{\epsilon} \right) J(p_1 \dots p_n, 0) + \int_0^1 \mathrm{d}x \, x^{-1-\epsilon} \, R_n(x) \, J(p_1 \dots p_{n+1}) \right\} . \tag{88}$$

In the inclusive case (calculation of the total cross section) we have $J \equiv 1$. The integration over x leads to the explicit $1/\epsilon$ poles which must cancel with the virtual part:

$$\int_0^1 \mathrm{d}x \, x^{-1-\epsilon} \, R_n(x) = -\frac{R_n(0)}{\epsilon} + \int_0^1 \mathrm{d}x \, x^{-\epsilon} \frac{R_n(x) - R_n(0)}{x} . \tag{89}$$

The cancellation of the poles between $\frac{V_n}{\epsilon}$ and $\frac{R_n(0)}{\epsilon}$ in the non-inclusive case will only work if

$$\lim_{p_{n+1} \to 0} J(p_1 \dots p_n, p_{n+1}) = J(p_1 \dots p_n, 0) . \tag{90}$$

This is a non-trivial condition for the definition of an observable, for example a jet algorithm, and is called *infrared safety*. The formulation above is taylored to the soft limit where all components of p_{n+1} go to zero, however, an analogous condition must hold if two momenta become collinear.

As mentioned above, the measurement function is also important if we define differential cross sections $\mathrm{d}\sigma/\mathrm{d}X$. In this case we have $J(p_1 \dots p_n) = \delta(X - \chi_n(p_i))$, where $\chi_n(p_i)$ is the definition of the observable, based on n partons. Again, infrared safety requires $\chi_{n+1} \to \chi_n$ if one of the p_i becomes soft or two of the momenta become collinear to each other.

Phase space integrals in D dimensions

To see how the cancellation works for *inclusive* quantities such as the total cross section, let us consider the real radiation contribution to $e^+e^- \to 2\,$jets at NLO in more detail. For this purpose we need phase space integrals in D dimensions.

The general formula for a $1 \to n$ particle phase space $\mathrm{d}\Phi_n$ with $Q \to p_1 \dots p_n$ is given by

$$\mathrm{d}\Phi_{1\to n} = (2\pi)^{n-D(n-1)} \Big[\prod_{j=1}^{n} d^D p_j \delta(p_j^2 - m_j^2)\Theta(E_j)\Big] \delta\Big(Q - \sum_{i=1}^{n} p_i\Big) \ . \tag{91}$$

In the following we will stick to the massless case $m_j = 0$. We use

$$d^D p_j \delta(p_j^2)\Theta(E_j) = dE_j d^{D-1}\vec{p}_j \delta(E_j^2 - \vec{p}_j^{\,2})\Theta(E_j) = \frac{1}{2E_j} d^{D-1}\vec{p}_j \Big|_{E_j = |\vec{p}_j|} \tag{92}$$

for $j = 1, \dots, n-1$ to arrive at

$$\mathrm{d}\Phi_{1\to n} = (2\pi)^{n-D(n-1)} 2^{1-n} \prod_{j=1}^{n-1} \frac{\mathrm{d}^{D-1}\vec{p}_j}{|\vec{p}_j|} \delta\Big((Q - \sum_{i=1}^{n-1} p_i)^2\Big) \ , \tag{93}$$

where we have used the last δ-function in Eq. (91) to eliminate p_n. We further use

$$\begin{aligned} &\frac{\mathrm{d}^{D-1}\vec{p}}{|\vec{p}|} f(|\vec{p}|) = \mathrm{d}\Omega_{D-2}\, \mathrm{d}|\vec{p}|\, |\vec{p}|^{D-3} f(|\vec{p}|) \ , \\ &\int \mathrm{d}\Omega_{D-2} = \int \mathrm{d}\Omega_{D-3} \int_0^\pi \mathrm{d}\theta (\sin\theta)^{D-3} = \int_0^\pi \mathrm{d}\theta_1 (\sin\theta_1)^{D-3} \int_0^\pi \mathrm{d}\theta_2 (\sin\theta_2)^{D-4} \dots \int_0^{2\pi} d\theta \ , \end{aligned} \tag{94}$$

to obtain

$$\mathrm{d}\Phi_{1\to n} = (2\pi)^{n-D(n-1)} 2^{1-n} \left(\prod_{j=1}^{n-1} \mathrm{d}\Omega_{D-1-j}\, \mathrm{d}|\vec{p}_j|\, |\vec{p}_j|^{D-3} \right) \delta\Big((Q - \sum_{i=1}^{n-1} p_i)^2\Big) \ . \tag{95}$$

Example* $1 \to 3$*:

For $n = 3$ one can choose a coordinate frame such that

$$\begin{aligned} &Q = (E, \vec{0}^{(D-1)}) \ , \quad p_1 = E_1\, (1, \vec{0}^{(D-2)}, 1) \ , \\ &p_2 = E_2\, (1, \vec{0}^{(D-3)}, \sin\theta, \cos\theta) \ , \quad p_3 = Q - p_2 - p_1 \ , \end{aligned} \tag{96}$$

leading to

$$\mathrm{d}\Phi_{1\to 3} = \frac{1}{4}(2\pi)^{3-2\,D}\, \mathrm{d}E_1 \mathrm{d}E_2 \mathrm{d}\theta_1\, (E_1 E_2 \sin\theta)^{D-3}\, \mathrm{d}\Omega_{D-2}\, \mathrm{d}\Omega_{D-3}$$

$$\Theta(E_1)\,\Theta(E_2)\,\Theta(E-E_1-E_2)\,\delta((Q-p_1-p_2)^2)\;. \tag{97}$$

In the following a parametrisation in terms of the Mandelstam variables $s_{ij} = 2\,p_i \cdot p_j$ will be useful, therefore we make the transformation $E_1, E_2, \theta \to s_{12}, s_{23}, s_{13}$. To work with dimensionless variables we define $y_1 = s_{12}/Q^2$, $y_2 = s_{13}/Q^2$, $y_3 = s_{23}/Q^2$ which leads to

$$\begin{aligned} \mathrm{d}\Phi_{1\to 3} = (2\pi)^{3-2\,D}\,2^{-1-D}(Q^2)^{D-3}\,\mathrm{d}\Omega_{D-2}\,\mathrm{d}\Omega_{D-3}\,\mathrm{d}y_1\,\mathrm{d}y_2\,\mathrm{d}y_3 \\ (y_1\,y_2\,y_3)^{D/2-2}\;\Theta(y_1)\,\Theta(y_2)\,\Theta(y_3)\,\delta(1-y_1-y_2-y_3)\;. \end{aligned} \tag{98}$$

Now we are in the position to calculate the full real radiation contribution. The matrix element (for one quark flavour with charge q_f) in the variables defined above, where p_3 is the gluon, is given by

$$|\mathcal{M}|^2_{\text{real}} = C_F e^2 q_f^2 g_s^2\,8\,(1-\epsilon)\left\{\frac{2}{y_2 y_3} + \frac{-2+(1-\epsilon)y_3}{y_2} + \frac{-2+(1-\epsilon)y_2}{y_3} - 2\epsilon\right\}\;. \tag{99}$$

In our variables, soft singularities mean $p_3 \to 0$ and therefore both y_2 and $y_3 \to 0$, while $p_3 \parallel p_1$ means $y_2 \to 0$ and $p_3 \parallel p_2$ means $y_3 \to 0$. Combined with the factors $(y_2\,y_3)^{D/2-2}$ from the phase space it is clear that the first term in the bracket of Eq. (99) will lead to a $1/\epsilon^2$ pole, coming from the region in phase space where soft and collinear limits coincide. The integrals can be expressed in terms of Euler Beta-functions and lead to the result quoted in Eq. (86).

4.2 Soft gluon emission

Soft gluon emission is very important in QCD. In contrast to the collinear case, soft gluons are insensitive to the spin of the partons. The only feature they are sensitive to is the colour charge.

To see this, consider the amplitude for the second row in Fig. 20, with momentum k and colour index a for the gluon, and momenta and colour indices p, i ($\bar{p}, j$) for the quark (antiquark). The amplitude for massless quarks is given by

$$\mathcal{M}^{a,\mu}_{ij} = t^a_{ij}\,g_s\,\mu^\epsilon \bar{u}(p)\;\not{\epsilon}(k)\frac{\not{p}+\not{k}}{(p+k)^2}\Gamma^\mu v(\bar{p}) - t^a_{ij}\,g_s\,\mu^\epsilon \bar{u}(p)\Gamma^\mu\frac{\not{\bar{p}}+\not{k}}{(\bar{p}+k)^2}\;\not{\epsilon}(k)v(\bar{p})\;, \tag{100}$$

where Γ^μ describes a general interaction vertex with the photon, in our case $\Gamma^\mu = \gamma^\mu$. Now we take the soft limit, which means that all components of k are much smaller than p and $\bar{p}$, thus neglecting factors of $\not{k}$ in the numerator and k^2 in the denominator. Using the Dirac equation leads to

$$\begin{aligned} \mathcal{M}^{a,\mu}_{ij,soft} &= g_s\,\mu^\epsilon\,t^a_{ij}\,\bar{u}(p)\,\Gamma^\mu\,v(\bar{p})\left(\frac{2\epsilon(k)\cdot p}{2p\cdot k} - \frac{2\epsilon(k)\cdot\bar{p}}{2\bar{p}\cdot k}\right) \\ &\equiv g_s\,\mu^\epsilon\,J^{a,\nu}_{ij}(k)\epsilon_\nu(k)\,\mathcal{M}^\mu_{Born}\;\;,\;\;\mathcal{M}^\mu_{Born} = \bar{u}(p)\Gamma^\mu v(\bar{p})\;. \end{aligned} \tag{101}$$

The amplitude factorises completely into the product of the Born amplitude and the *soft gluon current*

$$J^{a,\nu}_{ij}(k) = \sum_{r=p,\bar{p}} T^a_{ij}\,\frac{r^\nu}{r\cdot k}\;, \tag{102}$$

In our example $T^a_{ij} = t^a_{ij}$ for $r = p$ and $T^a_{ij} = -t^a_{ij}$ for $r = \bar{p}$. This type of factorisation actually holds for an arbitrary number of soft gluon emissions, and can be obtained using the "soft Feynman rules" shown in Fig. 22.

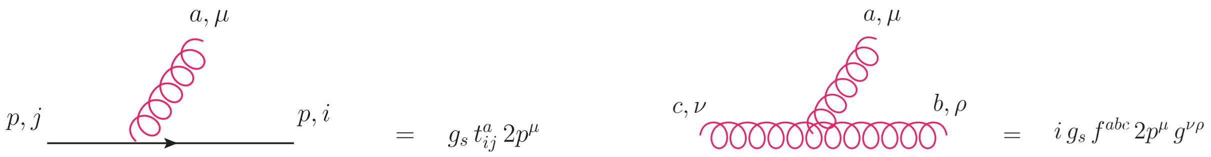

a, μ
p, j
p, i
$= \quad g_s\, t^a_{ij}\, 2p^\mu$
a, μ
c, ν
b, ρ
$= \quad i\, g_s\, f^{abc}\, 2p^\mu\, g^{\nu\rho}$

Fig. 22: The Feynman rules for gluon emission in the soft limit.

Following the standards set by Refs. [16,19], the soft gluon current is more conveniently expressed in terms of colour charge operators $\mathbf{T}_i$, where i now labels the *parton* i emitting a gluon (not its colour index).

The action of $\boldsymbol{T}_i$ onto the colour space is defined by

$$\langle a_1, \ldots, a_i, \ldots, a_m, a\,|\boldsymbol{T}_i|b_1, \ldots, b_i, \ldots, b_m\rangle = \delta_{a_1 b_1} \ldots . T^a_{a_i b_i} \ldots \delta_{a_m b_m} \quad , \tag{103}$$

where $T^a_{kl} \equiv t^a_{kl}$ ($SU(3)$ generator in the fundamental representation) if the emitting particle i is a quark. In the case of an emitting antiquark $T^a_{kl} \equiv \bar{t}^a_{kl} = -t^a_{lk}$. If the emitting particle i is a gluon, $T^a_{bc} \equiv -if_{abc}$ ($SU(3)$ generator in the adjoint representation).

Then we can write down the universal behaviour of the matrix element $\mathcal{M}(k, p_1, \ldots, p_m)$ in the limit where the momentum k of the gluon becomes soft. Denoting by a and $\varepsilon^\mu(k)$ the colour and the polarisation vector of the soft gluon, the matrix element fulfils the following factorisation formula:

$$\mathcal{M}^a(k, p_1, \ldots, p_m) \simeq g_s\, \mu^\epsilon \varepsilon^\mu(k)\; J^a_\mu(k)\; \mathcal{M}(p_1, \ldots, p_m)\ , \tag{104}$$

where $\mathcal{M}^a(p_1, \ldots, p_m)$ is obtained from the original matrix element by removing the soft gluon k. The factor $\boldsymbol{J}_\mu(k)$ is the soft-gluon current

$$\boldsymbol{J}^\mu(k) = \sum_{i=1}^{m} \boldsymbol{T}_i\, \frac{p_i^\mu}{p_i \cdot k}\ , \tag{105}$$

which depends on the momenta and colour charges of the hard partons in the matrix element on the right-hand side of Eq. (104). The symbol '$\simeq$' means that on the right-hand side we have neglected contributions that are less singular than $1/|k|$ in the soft limit $k \to 0$.

Squaring Eq. (104) and summing over the gluon polarisations leads to the *universal soft-gluon factorisation formula* at $\mathcal{O}(\alpha_s)$ for the squared amplitude [16]

$$|\mathcal{M}(k, p_1, \ldots, p_m)|^2 \simeq -g_s^2\, \mu^{2\epsilon}\, 2 \sum_{i,j=1}^{m} S_{ij}(k)\; |\mathcal{M}_{(i,j)}(p_1, \ldots, p_m)|^2 \quad , \tag{106}$$

where the factor

$$S_{ij}(p_s) = \frac{p_i \cdot p_j}{2\,(p_i \cdot p_s)\,(p_j \cdot p_s)} = \frac{s_{ij}}{s_{is}\, s_{js}} \tag{107}$$

is called the *Eikonal factor*. It can be generalised to the emission of n soft gluons and plays an important role in resummation.

The colour correlations produced by the emission of a soft gluon are taken into account by the square of the colour-correlated amplitude $|\mathcal{M}_{(i,j)}|^2$, given by

$$\begin{aligned}
&|\mathcal{M}_{(i,j)}(p_1,\ldots,p_m)|^2 \equiv \langle\, \mathcal{M}(p_1,\ldots,p_m)\,|\, \boldsymbol{T}_i\cdot\boldsymbol{T}_j \,|\, \mathcal{M}(p_1,\ldots,p_m)\rangle \\
&= \big(\,\mathcal{M}_{c_1..b_i...b_j...c_m}(p_1,\ldots,p_m)\big)^*\; T^a_{b_i d_i}\, T^a_{b_j d_j}\; \mathcal{M}_{c_1..d_i...d_j...c_m}(p_1,\ldots,p_m)\big)\;.
\end{aligned}$$

The angular brackets in the second line denote a basis in colour space.

4.3 Collinear singularities

Let us come back to the amplitude for the real radiation given in Eq. (100). In a frame where $p = E_p(1,\vec{0}^{(D-2)},1)$ and $k = k_0(1,\vec{0}^{(D-3)}\sin\theta,\cos\theta)$, the denominator $(p+k)^2$ is given by

$$(p+k)^2 = 2k_0E_p\,(1-\cos\theta) \quad \to 0 \quad \text{for} \left\{ \begin{array}{ll} k_0\to 0 & \text{(soft)} \\ \theta\to 0 & \text{(collinear)} \end{array}\right. \tag{108}$$

Note that if the quark line was massive, $p^2 = m^2$, we would have

$$(p+k)^2 - m^2 = 2k_0E_p\,(1-\beta\cos\theta)\,,\quad \beta = \sqrt{1-m^2/E_p^2}$$

and thus the collinear singularity would be absent. This is why collinear singularities are sometimes also called *mass singularities*, since the propagator can only develop a collinear divergence if the splitting partons are massless, while the soft singularity is present irrespective of the mass of the quark radiating a gluon.

The important point to remember is that in the collinear limit, we also have a form of factorisation, shown schematically in Fig. 23.

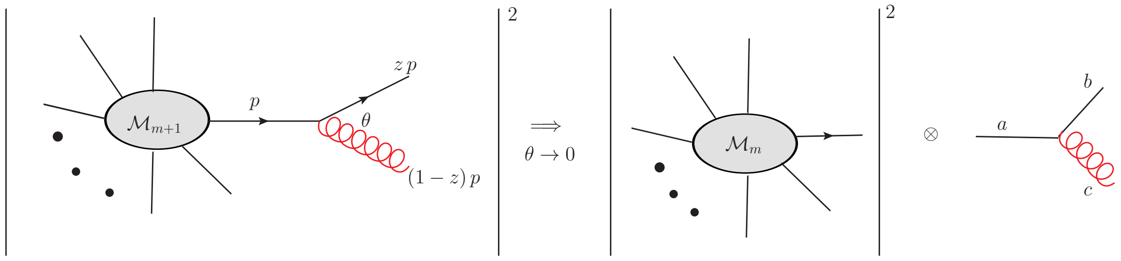


Fig. 23: Factorisation in the collinear limit.

The universal factorisation behaviour can be described as

$$|\mathcal{M}_{m+1}|^2\,\mathrm{d}\Phi_{m+1} \to |\mathcal{M}_m|^2\mathrm{d}\Phi_m\,\frac{\alpha_s}{2\pi}\,\frac{\mathrm{d}k_\perp^2}{k_\perp^2}\,\frac{\mathrm{d}\phi}{2\pi}\,\mathrm{d}z\,P_{a\to bc}(z)\;, \tag{109}$$

where we have used the so-called *Sudakov parametrisation*:

$$\begin{aligned} k^\mu &= (1-z)\,p^\mu + \beta\, n^\mu + k_\perp^\mu\;, \\ k^+ &= k\cdot n = (1-z)\,p\cdot n\;,\quad k^- = k\cdot p = \frac{k_\perp^2}{2(1-z)}\;, \end{aligned} \tag{110}$$

with n^μ being a light-like vector satisfying $p \cdot n \neq 0$ and $k_\perp \cdot n = 0$, and β being determined by the requirement that k must be light-like:

$$k^2 = 0 = 2(1-z)\,\beta\, p\cdot n - k_\perp^2 \Rightarrow \beta = \frac{k_\perp^2}{2\,p\cdot n\,(1-z)}\;. \tag{111}$$

The function $P_{a\to bc}(z)$ is the so-called *Altarelli–Parisi splitting function*, describing the splitting of parton a into partons b and c, and z is the momentum fraction of the original parton a taken away by parton b after emission of c. For example, consider collinear gluon emission off a quark:

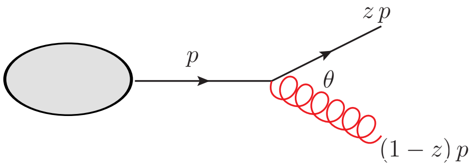


Fig. 24: Gluon emission leading to $P_{q\to qg}(z)$.

The corresponding Altarelli-Parisi splitting function for $z < 1$ is given by

$$P_{q\to qg}(z) = C_F\,\frac{1+z^2}{1-z}\;, \tag{112}$$

and is often just denoted as $P_{qq}(z)$. The other possible splitting functions have the following form:

$$\begin{aligned} P_{q\to gq}(z) &= C_F\,\frac{1+(1-z)^2}{z}\;,\quad P_{g\to q\bar q}(z) = T_R\,\left(z^2+(1-z)^2\right)\;, \\ P_{g\to gg}(z) &= C_A\left(z\,(1-z) + \frac{z}{1-z} + \frac{1-z}{z}\right)\;. \end{aligned} \tag{113}$$

We will come back to them later when we discuss parton distribution functions.

4.3.1 Jet cross sections and jet algorithms

Jets can be pictured as clusters of particles which are close to each other in phase space, resp. in the detector.

In Fig. 25 (left), a typical composition of the different types of particles making up jets, in terms of the fraction of the total jet energy carried by the particles, is shown for a simulated 2-jet event at the LHC. These jets are primarily composed of charged and neutral pions. Further, baryons and other types of mesons contribute a moderate fraction of the total jet energy, and small energy fractions of electrons and muons are also present, originating from heavy hadron decays. Until today, jets have been measured over a very large energy range at different colliders, see Fig. 25 (right).

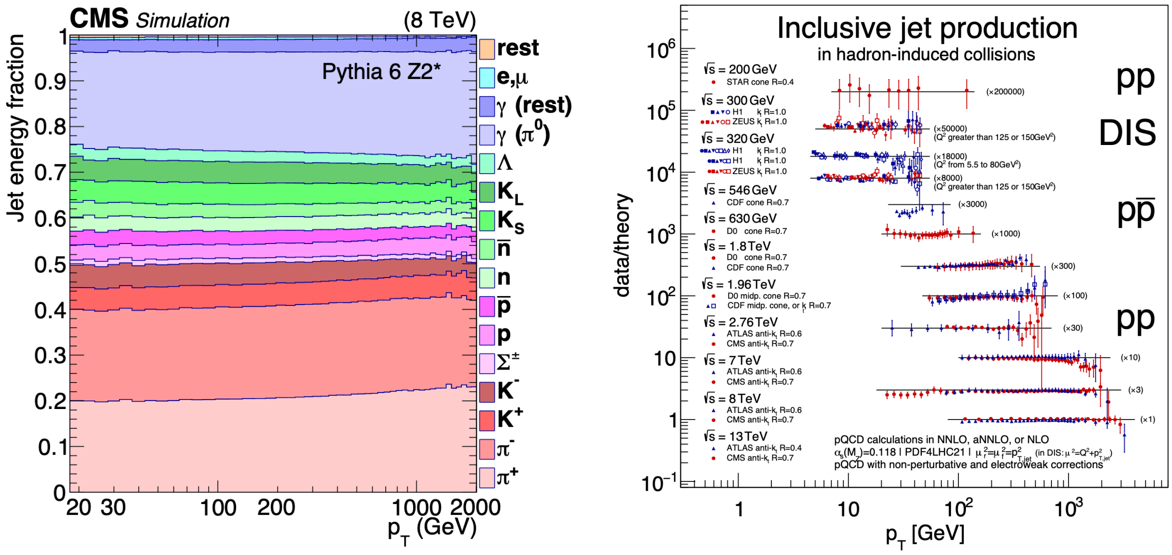


Fig. 25: Left: The fraction of the total jet energy carried by different types of particles forming jets (simulated LHC dijet events). Right: Data over theory predictions for inclusive jet production at different experiments and center-of-mass energies. *Figures from Ref. [48], B. Malaescu et al., D. Britzger et al.*

Sterman–Weinberg jet definition

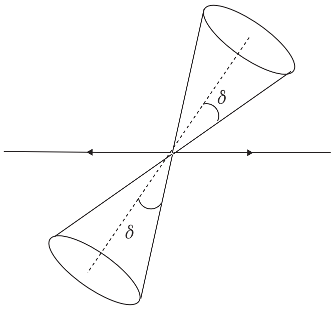


Fig. 26: Two jet cones according to the definition of Sterman and Weinberg.

Historically, one of the first suggestions to define jet cross sections was by Sterman and Weinberg [73]. In their definition, a final state is classified as two-jet-like if all but a fraction ε of the total available energy E is contained in two cones of opening angle δ. The two-jet cross section is then obtained by integrating the matrix elements for the various quark and gluon final states over the appropriate region of phase space determined by ε and δ.

Of course the two-jet cross section depends on the values for ε and δ. If they are very large, even extra radiation at a relatively large angle $\theta < \delta$ will be "clustered" into the jet cone and almost all events will be classified as 2-jet events. If they are very small, the 2-jet cross section starts to diverge, because "one parton" is not an observable, it cannot be distingushed from "one parton plus soft and/or collinear radiation".

More modern jet algorithms

The Sterman–Weinberg jet definition based on cones is not very practical to analyse multijet final states. A better alternative is for example the following:

1. starting from n particles, for all pairs i and j calculate $(p_i + p_j)^2$.
2. If $\min(p_i + p_j)^2 < y_{\text{cut}}\, Q^2$ then define a new "pseudo-particle" $p_J = p_i + p_j$, which decreases $n \to n-1$. Q is the center-of-mass energy in e^+e^- collisions, or a typical hard scattering energy in hadronic collisions, and y_{cut} is the jet resolution parameter.
3. if $n = 1$, stop, else repeat the step above.

After this algorithm, all partons are clustered into jets. This simple algorithm is sometimes called JADE-algorithm because it has been used first at the JADE experiment at PETRA (DESY). With this definition one finds at $\mathcal{O}(\alpha_s)$:

$$\sigma^{2\,jet} = \sigma_0 \left(1 - C_F\, \frac{\alpha_s}{\pi} \left[\ln^2 y_{\text{cut}} + \frac{3}{2} \ln y_{\text{cut}} + \text{ finite}\right]\right) \; . \tag{114}$$

Algorithms which are particularly useful for hadronic initial states are e.g. the so-called Durham-k_T algorithm [10] or the anti-k_T algorithm [14]. Both algorithms are based on a distance measure

$$d_{ij} = \min(p_{T,i}^{2p}, p_{T,j}^{2p}) \frac{\Delta R_{ij}^2}{R^2} \; , \tag{115}$$

where R is a radius parameter, $\Delta R_{ij}^2 = \Delta y_{ij}^2 + \Delta\phi_{ij}^2$ is the distance in rapidity and azimuthal angle between particles i and j, and the parameter p is 1 for the k_T algorithm, 0 for the Cambridge–Aachen [33] algorithm and -1 for the anti-k_T algorithm. The distance d_{ij} is calculated for all combinations of pairs of particles. The pair with the lowest d_{ij} is replaced by a pseudo-particle whose four-momentum is given by the sum of the four-momenta of particles i and j. Summing the 4-momenta to form the pseudo-particle is also called "E-recombination scheme". Note that the combined 4-momentum is not light-like anymore. The clustering procedure is repeated as long as pairs with invariant mass fraction below a predefined resolution parameter y_{cut} are found. Once the clustering is terminated, the remaining (pseudo-)particles are the jets. Figures 27 and 28 illustrate how the jet areas depend on the jet algorithm and the jet radius R.

It is evident that a large value of y_{cut} will ultimately result in the clustering all particles into only two jets, while higher jet multiplicities will become more and more frequent as y_{cut} is lowered. In experimental jet measurements, one therefore studies the jet rates (n-jet cross sections normalised to the total hadronic cross section) as function of the jet resolution parameter y_{cut}. Figure 29 (left) shows the jet rates as a function of y_{cut}, compared to ALEPH data. Figure 29 (right) shows corrections up to NNLO to the 3-jet rate as a function of y_{cut}. Note that in this figure, for small values of y_{cut}, the 3-jet rate at LO diverges (green band) because it shows a partonic calculation: only three partons are present at LO and therefore there is no room for extra radiation. As an isolated parton is not an observable, the cross section diverges in this limit. At higher orders, this situation is somewhat cured by extra radiation being allowed, however resummation or parton showering would be needed to achieve a better description of the very low y_{cut} region.

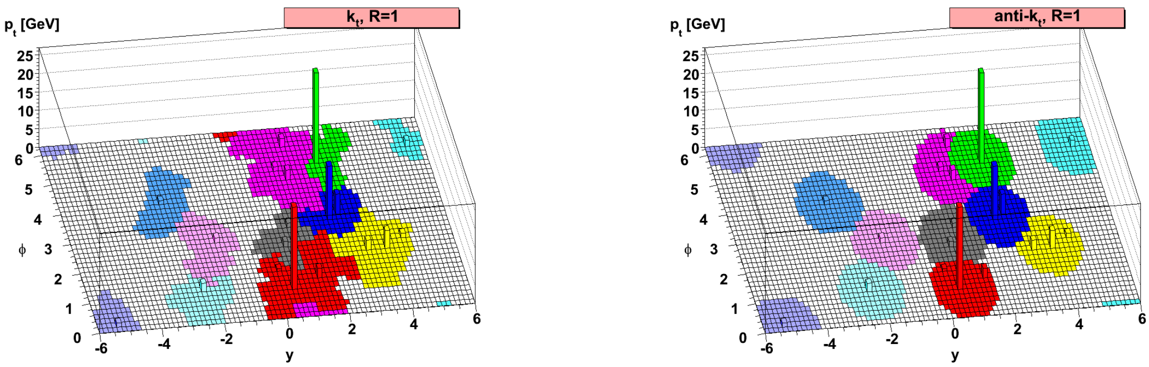


Fig. 27: Jet areas as a result of (a) the Durham-k_T algorithm, (b) the anti-k_T algorithm. Figures from Ref. [14].

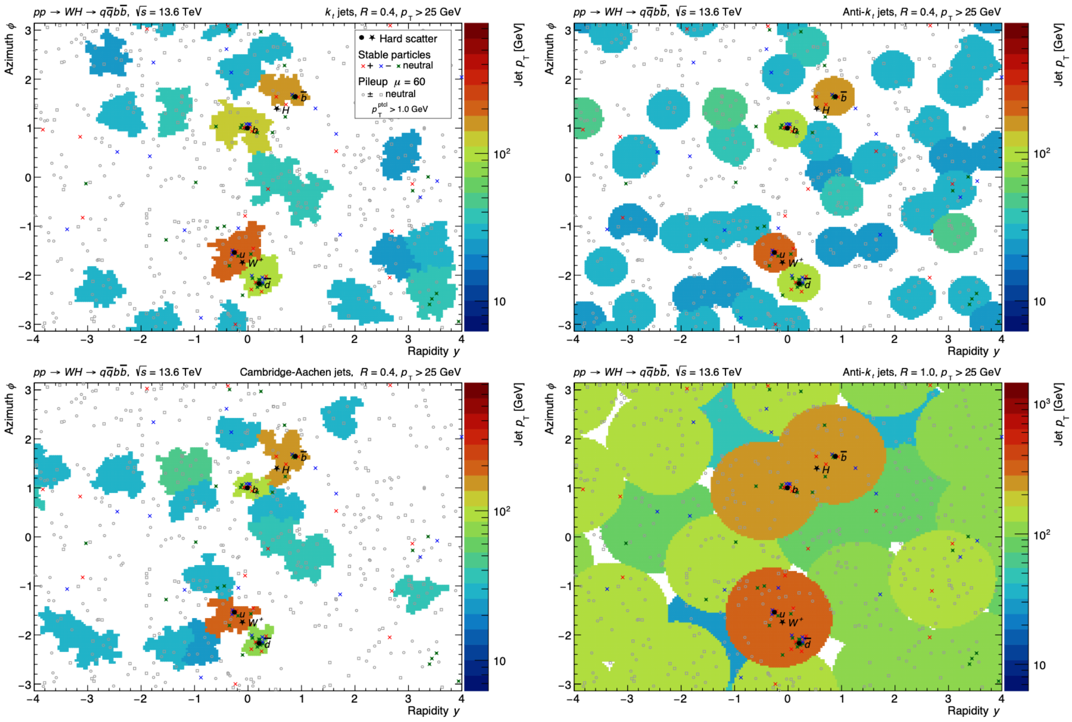


Fig. 28: Jet areas as a result of different jet algorithms and jet radii R. Figure from Ref. [48].

At the LHC, the most commonly used jet algorithm is the *anti-k_T algorithm* [14]. Of course it is very important that jet algorithms are infrared safe. The standard tool for jet identification in simulation programs is `fastjet`, see `http://fastjet.fr/`. More details about jet algorithms can be found in Ref. [71].

4.3.2 Jet substructure

The investigation of jet substructure is relatively recent and has developed into an essential tool for the LHC physics program. Information about the jet substructure is very useful to disentangle different kinds of jets, such as separating quark-initiated jets from gluon-initiated jets or isolating jets from boosted W, Z, H- or top-quark-decays from the background of quark and gluon-initiated jets.

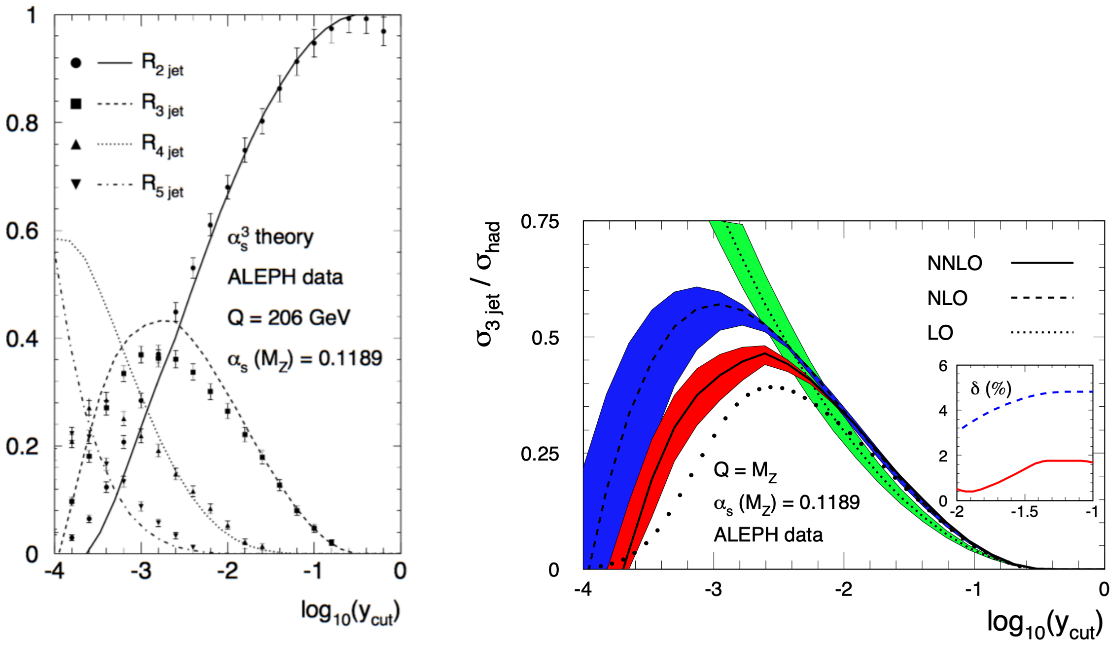


Fig. 29: Left: Jet rates as a function of the jet resolution parameter y_{cut} [51]. Right: higher-order corrections to the 3-jet rate [41].

Jet substructure tools aim to get information about the internal kinematic properties of a high-p_T jet. Important quantities in this context are the invariant jet mass M^2_{jet}, defined as

$$M^2_{\text{jet}} = \left(\sum_{i \in \text{jet}} p_i \right)^2 , \tag{116}$$

where the p_i denote the 4-momenta of the jet constituents, and the generalised jet angularities λ_β, defined as [64]

$$\lambda_\beta = \sum_{i \in \text{jet}} z_i \left(\frac{\Delta R_{i,\text{jet}}}{R} \right)^\beta , \tag{117}$$

where z_i is the jet transverse momentum fraction carried by the constituent i and $\Delta R_{i,\text{jet}}$ is its distance to the jet axis in the $(y - \phi)$–plane:

$$z_i = \frac{p_{t,i}}{\sum_{j \in \text{jet}} p_{t,j}} \qquad \text{and} \qquad \Delta R^2_{i,\text{jet}} = (y_i - y_{\text{jet}})^2 + (\phi_i - \phi_{\text{jet}})^2 . \tag{118}$$

Gluon-initiated jets in general have larger angularities than quark-initiated jets. A large variety of methods for jet substructure have been proposed over the last ten years, for more details we refer to Ref. [64].

While the jet mass and other jet substructure indicators provide a solid baseline, modern jet substructure classifiers make use of machine learning techniques to maximally discriminate between different possible jet origin interpretations.

4.3.3 Event shapes

Of course, jets are not the only observables one can define based on hadronic tracks in the detector. Other very useful observables are so-called *event-shape* observables, for example *thrust*, which describes how "pencil-like" an event looks. Thrust T is defined by

$$T = \max_{\vec{n}} \frac{\sum_{i=1}^{m} |\vec{p}_i \cdot \vec{n}|}{\sum_{i=1}^{m} |\vec{p}_i|} \,, \tag{119}$$

where $\vec{n}$ is a three-vector (the direction of the thrust axis) such that T is maximal. The particle three-momenta $\vec{p}_i$ are defined in the centre-of-mass frame. Therefore, the above definition only holds for lepton colliders where the partonic centre-of-mass energy is fixed. At hadron colliders, the definition of event shapes such as thrust is still possible, but in this case it is based on transverse momenta. T is an example of a measurement function $J(p_1, \ldots, p_m)$. It is infrared safe because neither $p_j \to 0$, nor replacing p_i with $zp_i + (1-z)p_i$ change T. Figure 30 shows the collinear and soft regions in a Dalitz-plot, where x_i denote the energy fractions, defined by

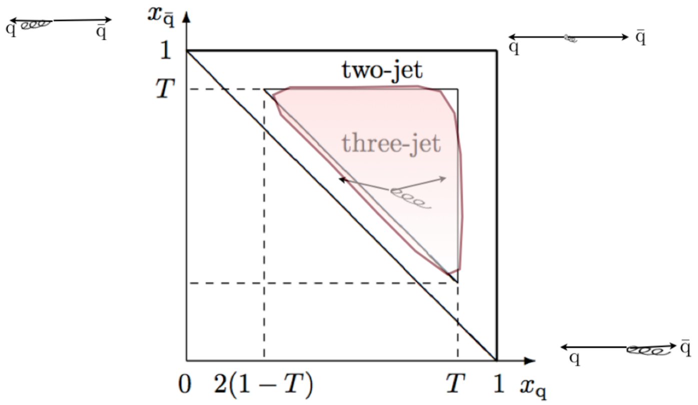


Fig. 30: Dalitz-plot showing the allowed 2-jet and 3-jet regions and thrust values. Figure from Ref. [30].

$$x_q = 2\frac{E_q}{\sqrt{s}}\,, \quad x_{\bar{q}} = 2\frac{E_{\bar{q}}}{\sqrt{s}}\,, \quad x_g = 2\frac{E_g}{\sqrt{s}}\,, \quad x_q + x_{\bar{q}} + x_g = 2\,. \tag{120}$$

At leading order it is possible to perform the phase space integrations analytically, to obtain

$$\frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}T} = C_F \frac{\alpha_s}{2\pi}\left[\frac{2\left(3T^2 - 3T + 2\right)}{T(1-T)} \ln\left(\frac{2T-1}{1-T}\right) - 3(3T-2)\frac{2-T}{1-T}\right]. \tag{121}$$

We see that the perturbative prediction for the thrust distribution becomes singular as $T \to 1$. In addition to the factor of $1-T$ in the denominator, there is also a logarithmic divergence $\sim \ln(1-T)$. The latter is characteristic for event shape distributions. For an event shape Y with $Y \to 0$ in the two-jet limit (so for example $Y = 1-T$), the behaviour at n-th order in perturbation theory is [17] $\frac{1}{\sigma}\frac{d\sigma^{(n)}}{dY} \simeq \alpha_s^n \frac{1}{Y} \ln^{2n-1}(\frac{1}{Y})$. These logarithms spoil the convergence of the perturbative series and should be "resummed" if we want to make reliable prediction near the phase-space region where $Y \to 0$. Summing the leading logarithms to all orders leads to an exponential function providing a "damping factor", the so-called Sudakov-factor.

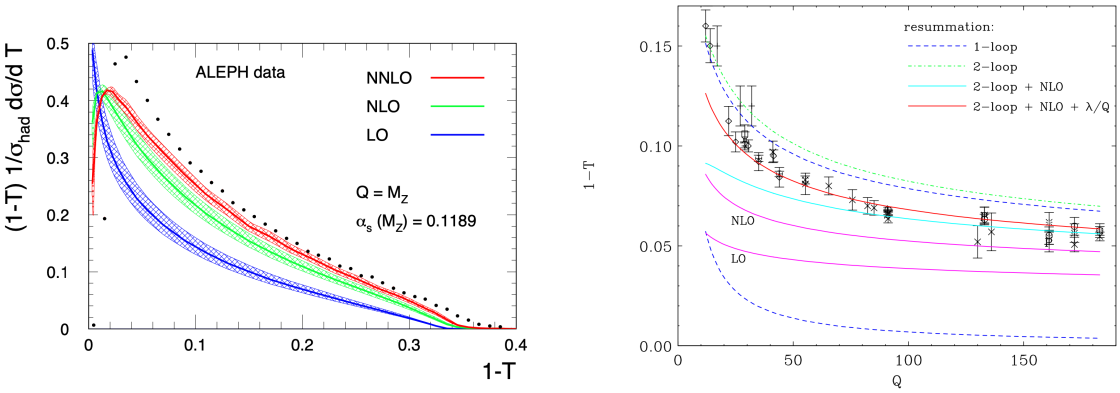


Fig. 31: Left: The thrust distribution up to NNLO in QCD, compared to ALEPH data. Figure from Ref. [40]. Right: The thrust distribution up to NLO in QCD and including resummation and power corrections, compared to data. Figure from Ref. [39].

Figure 31 (left) shows the thrust distribution up to NNLO precision in QCD. This is an observable where both resummation and power corrections $\sim \lambda/Q$ need to be included to describe the data well over the whole kinematic range, as can be seen from Fig. 31 (right).

5 PDFs and parton evolution

5.1 Deeply inelastic scattering

In the discussion of the cancellation of IR singularities we have only considered leptons in the initial state (e^+e^- annihilation). Now we consider the case where we have an electron–proton collider, like for example HERA, which operated at DESY until 2007 and offered unique opportunities to study the proton structure. We consider the scattering of electrons off the proton by photon exchange, as depicted in Fig. 32, in a kinematic regime where the squared momentum transfer Q^2 is large compared to the proton mass squared ($M \sim 1\,\mathrm{GeV}$), so we consider deeply inelastic scattering (DIS), $e(k)+p(P) \to e(k')+X$, where P is the momentum of the proton. The relations between the involved momenta and some useful

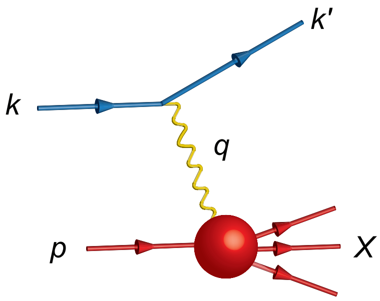


Fig. 32: Deeply inelastic scattering. Figure from Ref. [2].

kinematic variables are

$$s = (P+k)^2 \quad \text{[cms energy]}^2$$

$$\begin{aligned}
q^\mu &= k^\mu - k'^\mu \quad \text{[momentum transfer]} \\
Q^2 &= -q^2 = 2MExy \\
x &= \frac{Q^2}{2P\cdot q} \quad \text{[scaling variable]} \\
y &= \frac{P\cdot q}{P\cdot k} = 1 - \frac{E'}{E} \quad \text{[relative energy loss]}\ .
\end{aligned} \tag{122}$$

The cross section can be written as

$$\int \mathrm{d}\sigma = \sum_X \frac{1}{4ME} \int \mathrm{d}\Phi\, \frac{1}{4} \sum_{\text{spins}} |\mathcal{M}|^2\ . \tag{123}$$

We can factorise the squared matrix element into a leptonic tensor $L^{\mu\nu}$ and a hadronic tensor $H_{\mu\nu}$, and also factorise the phase space,

$$\mathrm{d}\Phi = \frac{\mathrm{d}^3k'}{(2\pi)^3 2E'}\,\mathrm{d}\Phi_X\ , \quad \frac{1}{4}\sum_{\text{spins}} |\mathcal{M}|^2 = \frac{e^4}{Q^4} L^{\mu\nu} H_{\mu\nu}\ . \tag{124}$$

Then the hadronic part of the cross section can be described by the dimensionless Lorentz tensor $W_{\mu\nu} = \frac{1}{8\pi}\sum_X \int \mathrm{d}\Phi_X H_{\mu\nu}$. As it depends only on two momenta P^μ and q^μ, the most general gauge- and Lorentz-invariant expression must be of the form

$$\begin{aligned}
W_{\mu\nu}(P,q) &= \left(-g_{\mu\nu} + \frac{q_\mu q_\nu}{q^2}\right) W_1(x,Q^2) \\
&+ \left(P_\mu - q_\mu \frac{P\cdot q}{q^2}\right)\left(P_\nu - q_\nu \frac{P\cdot q}{q^2}\right) \frac{W_2(x,Q^2)}{P\cdot q}\ ,
\end{aligned} \tag{125}$$

where the structure functions $W_i(x,Q^2)$ are dimensionless functions of the scaling variable x and the momentum transfer Q^2.

For the leptonic part we use the relations $E' = (1-y)E$, $\cos\theta = 1 - \frac{xyM}{(1-y)E}$ to change variables to x and the relative energy loss y,

$$\frac{\mathrm{d}^3k'}{(2\pi)^3 2E'} = \frac{\mathrm{d}\phi}{2\pi}\frac{E'}{8\pi^2}\,\mathrm{d}E'\,\mathrm{d}\cos\theta = \frac{\mathrm{d}\phi}{2\pi}\,\frac{yME}{8\pi^2}\,\mathrm{d}y\,\mathrm{d}x\,,$$

and compute the trace $L^{\mu\nu} = \frac{1}{2}Tr[\not{k}\gamma^\mu\ \not{k}'\gamma^\nu] = k^\mu k'^\nu + k^\nu k'^\mu - g^{\mu\nu} k\cdot k'$. Then the differential cross section in x and y is obtained from Eq. (123) as

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}x\,\mathrm{d}y} = \frac{4\pi\alpha^2}{y\,Q^2}\left[y^2 W_1(x,Q^2) + \left(\frac{1-y}{x} - xy\frac{M^2}{Q^2}\right) W_2(x,Q^2)\right]\ .$$

In the *scaling limit*, defined by $Q^2 \to \infty$ with x fixed, we use $W_1 \to -F_1, W_2 \to F_2$, neglect the term $\sim M^2/Q^2$ and obtain

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}x\,\mathrm{d}y} = \frac{4\pi\alpha^2}{y\,Q^2}\left[\left(1+(1-y)^2\right)F_1 + \frac{1-y}{x}\left(F_2 - 2xF_1\right)\right]\ . \tag{126}$$

The functions F_1 and F_2 are called *structure functions*, where the combination $F_L = F_2 - 2xF_1$ is also called the longitudinal structure function because it is related to the absorption of a longitudinally polarised virtual photon. They were first measured by the SLAC–MIT experiment (USA) in 1970, and have been measured very accurately at the HERA collider. A very interesting feature is the fact that, in the scaling limit, we observe that $2xF_1 \to F_2$ and that F_2 becomes independent of Q^2, $F_2(x, Q^2) \to F_2(x)$, a feature which is often called *Bjorken scaling*. Furthermore, the *Callan–Gross* relation $F_2(x) = 2x\,F_1(x)$ can be derived from first principles under the assumption that the photon scatters off point-like spin-1/2 particles. These observations were very important to establish the quark model. How the scaling looks in experiment is shown in Fig. 33, where we see that scaling violations are present, increasing at small x. This can be explained by considering corrections to the naïve quark model.

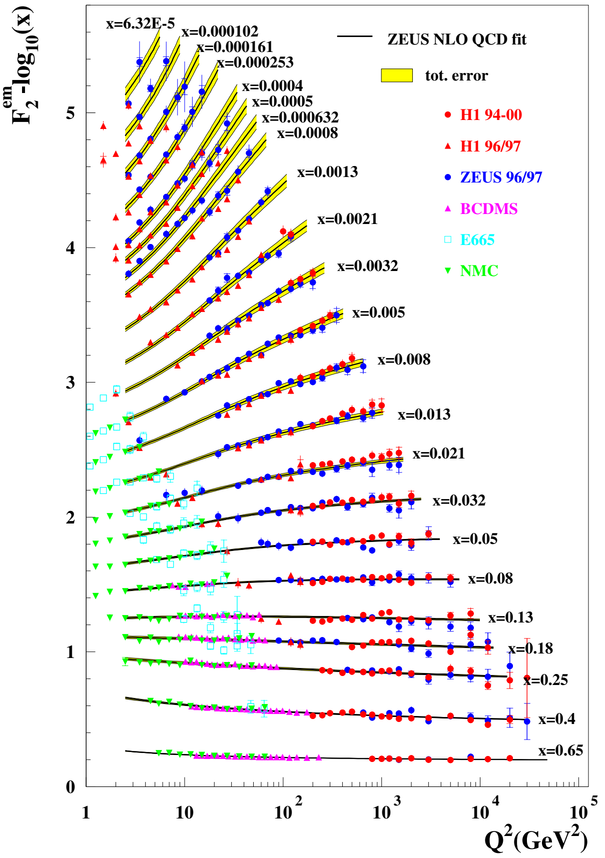


Fig. 33: The structure function F_2 for different values of Q^2. Figure from Ref. [59].

5.2 Proton structure in the parton model

Now let us assume the proton consists of free quarks and the lepton exchanges a hard virtual photon with one of those quarks. The struck quark carries a momentum p^μ, which is a fraction of the proton momentum, $p^\mu = \xi P^\mu$, so we consider the process $e(k) + q(p) \to e(k') + q(p')$. The corresponding cross section is

$$\hat{\sigma} = \frac{1}{2\hat{s}} \int \mathrm{d}\Phi_2 \, \frac{1}{4} \sum_{\text{spins}} |\mathcal{M}|^2 \, . \tag{127}$$

with $\hat{s} = (p+k)^2$. The "hat" indicates that we consider the partonic cross section. The squared matrix element is proportional to the product of the leptonic tensor $L^{\mu\nu}$ and a quark tensor $Q_{\mu\nu}$, with $Q_{\mu\nu} = \frac{1}{2} Tr[\not{p}\gamma^\mu \not{p}'\gamma^\nu] = p^\mu p'^\nu + p^\nu p'^\mu - g^{\mu\nu} p \cdot p'$, leading to $L^{\mu\nu} Q_{\mu\nu} = 2(\hat{s}^2 + \hat{u}^2)$, where $\hat{u} = (p-k')^2 = -2p \cdot k'$. As $y = Q^2/\hat{s}$ we can derive, using $\hat{u}^2 = (1-y)^2 \hat{s}^2$,

$$\frac{1}{4} \sum_{\text{spins}} |\mathcal{M}|^2 = \frac{e_q^2 e^4}{Q^4} L^{\mu\nu} Q_{\mu\nu} = 2 e_q^2 e^4 \frac{\hat{s}^2}{Q^4} \left(1 + (1-y)^2\right) \, . \tag{128}$$

Using $p'^2 = 2p \cdot q - Q^2 = Q^2(\xi/x - 1)$, the two-particle phase space (in 4 dimensions) can be written as

$$\mathrm{d}\Phi_2 = \frac{\mathrm{d}^3 k'}{(2\pi)^3 2E'} \frac{\mathrm{d}^4 p'}{(2\pi)^3} \, \delta\!\left(p'^2\right) (2\pi)^4 \, \delta^{(4)}(k + p - k' - p') = \frac{\mathrm{d}\phi}{(4\pi)^2} \, \mathrm{d}y \, \mathrm{d}x \, \delta(\xi - x) \, . \tag{129}$$

The differential cross section in x and y for one quark flavour is then given by

$$\frac{\mathrm{d}^2 \hat{\sigma}}{\mathrm{d}x \, \mathrm{d}y} = \frac{4\pi\alpha^2}{yQ^2} \left[1 + (1-y)^2\right] \frac{1}{2} e_q^2 \delta(\xi - x) \, . \tag{130}$$

Comparing Eqs. (126) and (130), we find the parton model predictions

$$\hat{F}_1(x) \propto e_q^2 \delta(\xi - x) \, , \qquad F_2 - 2xF_1 = 0 \, . \tag{131}$$

Therefore we found that the *Callan–Gross* relations follow directly from the assumption that a quark contained in the proton is responsible for the hard scattering. Thus the structure functions probe the quark constituents of the proton with $\xi = x$. However, this prediction cannot be the end of the story because experimentally, we observe that F_2 does depend on Q^2, as can be seen from Fig. 33, even though the dependence is not strong.

To see how the Q^2 dependence comes in, let us define the following:

$f_i(\xi)\mathrm{d}\xi$ is the probability to find a parton $(q, \bar{q}, g)$ with flavour i in the proton, carrying a momentum fraction of the proton between ξ and $\xi + \delta\xi$.

The function $f_i(\xi)$ is called *parton distribution function (PDF).*

Using the relations $\mathrm{d}y = \mathrm{d}Q^2/\hat{s}$ and $\delta(\xi - x) = \frac{1}{\xi}\delta\left(1 - \frac{x}{\xi}\right)$, we can write the full cross section as a combination of the PDF and the partonic cross section (130),

$$\frac{\mathrm{d}^2 \sigma}{\mathrm{d}x \, \mathrm{d}Q^2} = \int_x^1 \frac{\mathrm{d}\xi}{\xi} \sum_i f_i(\xi) \frac{\mathrm{d}^2 \hat{\sigma}}{\mathrm{d}x \, \mathrm{d}Q^2} \left(\frac{x}{\xi}, Q^2\right) \, . \tag{132}$$

This means that the cross section is a convolution of a long-distance component, the parton distribution function $f_i(\xi)$ for a parton of type i, and a short-distance component, the partonic hard scattering cross section $\hat{\sigma}$. This *factorisation* (up to power corrections), mentioned already in Section 2.1, can be proven rigorously in DIS using operator product expansion, and less rigorously in hadron–hadron collisions.

According to eqs. (126) and (132), we find in the naïve parton model

$$F_2(x) = 2xF_1(x) = \sum_i \int_0^1 \mathrm{d}\xi \, f_i(\xi)\, x\, e_{q_i}^2 \, \delta(x-\xi) = x \sum_i e_{q_i}^2 \, f_i(x) \; . \tag{133}$$

For a proton probed at a scale Q, we expect it to consist mostly of uud. Writing $f_i(x) = u(x), d(x)$ etc. for $i = u, d, \ldots$ we have in the naïve parton model

$$F_2^{\text{proton}}(x) = x\Big[\frac{4}{9}\big(u(x)+\bar{u}(x)\big) + \frac{1}{9}\big(d(x)+\bar{d}(x)\big)\Big] \; . \tag{134}$$

If we define the so-called "valence quarks" $u_v(x)u_v(x)d_v(x)$,

$$u(x) = u_v(x) + \bar{u}(x) \; , \; d(x) = d_v(x) + \bar{d}(x) \; , \; s(x) = \bar{s}(x) \; ,$$

we expect the "sum rules"

$$\int_0^1 \mathrm{d}x \, u_v(x) = 2 \; , \quad \int_0^1 \mathrm{d}x \, d_v(x) = 1 \; , \quad \int_0^1 \mathrm{d}x \, (s(x) - \bar{s}(x)) = 0 \; . \tag{135}$$

Figure 34 illustrates that, the smaller x and the larger Q^2, the more the "sea quarks" and gluons in the proton are probed. In fact, it turns out that $\sum_{i=q,\bar{q}} \int_0^1 \mathrm{d}x \, x f_{i/p}(x) \simeq 0.5$, so quarks carry only about half of the momentum of the proton. Therefore the other half is carried by gluons; the naïve parton model is clearly not sufficient to describe the proton.

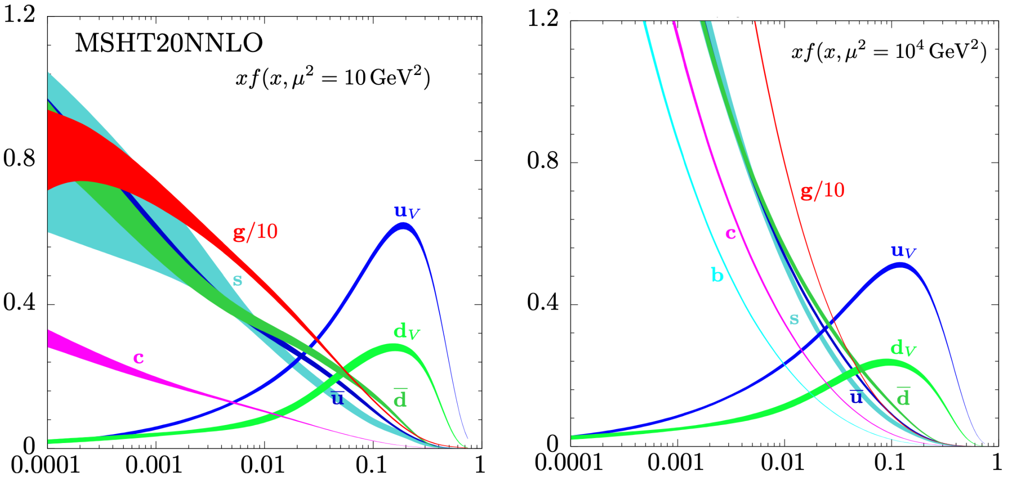
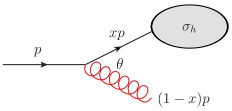


Fig. 34: Parton distribution functions in the proton as a function of x. Source: Particle Data Group, section *structure functions* [77].

5.3 Proton structure in perturbative QCD

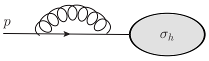
Investigating what happens in the "QCD-improved" parton model, we will encounter again IR singularities and splitting functions. Let us denote the hard scattering cross section by σ_h. For final-state radiation, we found that the IR singularities due to soft and collinear configurations cancel against IR divergences in the virtual corrections for infrared safe quantities.

If there is a coloured parton in the *initial* state, the splitting may occur *before* the hard scattering, such that the momentum of the parton that enters the hard process is reduced to xp^μ.



$$\mathrm{d}\sigma_{h+g}(p) \simeq \sigma_h(xp)\, 2C_F \frac{\alpha_s}{\pi}\frac{\mathrm{d}E}{E}\frac{\mathrm{d}\theta}{\theta} \to \sigma_h(xp)\, C_F\frac{\alpha_s}{\pi}\,\mathrm{d}x\,(1-x)^{-1-\epsilon}\,\mathrm{d}k_\perp^2\,(k_\perp^2)^{-1-\epsilon}\,.$$

Integrating over x and $k_\perp$ we find a soft divergence for $x \to 1$ and a collinear divergence for $k_\perp \to 0$. The corresponding $1/\epsilon$ poles multiply $\sigma_h(xp)$, while in the virtual correction the poles multiply $\sigma_h(p)$, irrespective whether the IR divergence is in the initial or final state:



$$\mathrm{d}\sigma_V \simeq -\sigma_h(p)\, C_F\frac{\alpha_s}{\pi}\,\mathrm{d}x\,(1-x)^{-1-\epsilon}\,\mathrm{d}k_\perp^2\,(k_\perp^2)^{-1-\epsilon}\,.$$

Summing up the real and virtual corrections we remain with an uncanceled collinear singularity from the factor $(k_\perp^2)^{-1-\epsilon}$,

$$\mathrm{d}(\sigma_{h+g}+\sigma_V) \simeq C_F\frac{\alpha_s}{\pi}\,\mathrm{d}k_\perp^2\,(k_\perp^2)^{-1-\epsilon}\mathrm{d}x\,\underbrace{(1-x)^{-1-\epsilon}[\sigma_h(xp)-\sigma_h(p)]}_{\text{finite}}\,, \tag{136}$$

Note that the soft singularity for $x \to 1$ is regulated in the sum of real and virtual parts. The uncanceled collinear singularity in the initial state however remains. Fortunately its form is universal, i.e. independent of the details of the hard scattering process, only dependent on the type of parton splittings. Therefore we can also eliminate it in a universal way: It is absorbed into "bare" parton densities, $f_i^{(0)}(x)$, such that the measured parton densities are the "renormalised" ones. This procedure is very similar to the renormalisation of UV divergences and introduces a scale μ_f, the *factorisation scale*, into the parton densities. Let us see how this works for the structure function F_2. We first consider the *partonic* structure functions $\hat{F}_{2,q}$, $\hat{F}_{2,g}$, where the subscript q indicates that a quark is coming out of the proton, analogous for a gluon g. Note that a gluon coming from the proton does not interact with a photon, therefore the gluonic contribution is zero at leading order, but it will appear at order α_s because the gluon can split into a $q\bar{q}$ pair and then one of the quarks interacts with the photon. Hence we have

$$\hat{F}_{2,q}(x) = \left.\frac{\mathrm{d}^2\hat{\sigma}}{\mathrm{d}x\,\mathrm{d}Q^2}\right|_{F_2} = e_q^2 x\left[\delta(1-x) + \frac{\alpha_s}{4\pi}\left(-\left(\frac{Q^2}{\mu^2}\right)^{-\epsilon}\frac{1}{\epsilon}\,P_{q\to qg}(x) + C_2^q(x)\right)\right]\,, \tag{137}$$

$$\hat{F}_{2,g}(x) = \left.\frac{\mathrm{d}^2\hat{\sigma}}{\mathrm{d}x\,\mathrm{d}Q^2}\right|_{F_2} = \sum_q e_q^2 x\left[0 + \frac{\alpha_s}{4\pi}\left(-\left(\frac{Q^2}{\mu^2}\right)^{-\epsilon}\frac{1}{\epsilon}\,P_{g\to q\bar{q}}(x) + C_2^g(x)\right)\right]\,, \tag{138}$$

where $P_{j\to ik}(x)$ is the Altarelli–Parisi splitting function (regularised at $x=1$) which we already encountered when discussing collinear singularities. It denotes the probability that a parton j splits collinearly into partons i and k, with i carrying a momentum fraction x of the original parton j. Note that the type of parton k is fixed by i and j. Therefore i and j are sufficient to label the splitting functions. For the labelling different conventions are in use, they are summarised in Table 2. $C_2(x)$ is the remaining finite term, sometimes called coefficient function. The partonic function $\hat{F}_2$ is not measurable, only the proton structure function F_2 is physical. Therefore we have to form the convolution of the partonic part with the parton distribution functions.

$P_{ij}(x)$	$P_{j\to ik}(x)$	$P_{i/j}(x)$
$P_{qq}(x)$	$P_{q\to qg}(x)$	$P_{q/q}(x)$
$P_{gq}(x)$	$P_{q\to gq}(x)$	$P_{g/q}(x)$
$P_{qg}(x)$	$P_{g\to q\bar{q}}(x)$	$P_{q/g}(x)$
$P_{gg}(x)$	$P_{g\to gg}(x)$	$P_{g/g}(x)$

Table 2: Translation between different conventions for the labelling of the splitting functions, see also Fig. 35.

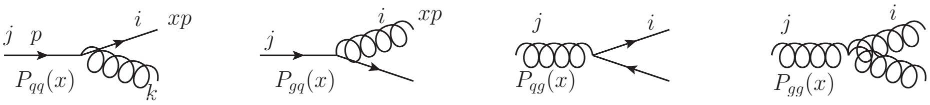


Fig. 35: Splitting functions with corresponding labelling.

$$\begin{aligned} F_{2,q}(x,Q^2) = x\sum_i e_{q_i}^2 \Bigg[& f_i^{(0)}(x) \\ & + \frac{\alpha_s}{2\pi}\int_x^1 \frac{\mathrm{d}\xi}{\xi} f_i^{(0)}(\xi)\left(-\left(\frac{Q^2}{\mu^2}\right)^{-\epsilon}\frac{1}{\epsilon}\,P_{q\to qg}\left(\frac{x}{\xi}\right) + C_2^q\left(\frac{x}{\xi}\right)\right)\Bigg] . \end{aligned} \tag{139}$$

Now we absorb the singularity into the parton distribution function by the definition

$$f_i(x,\mu_f^2) = f_i^{(0)}(x) + \frac{\alpha_s}{2\pi}\int_x^1 \frac{\mathrm{d}\xi}{\xi}\left\{ f_i^{(0)}(\xi)\left[-\frac{1}{\epsilon}\left(\frac{\mu_f^2}{\mu^2}\right)^{-\epsilon} P_{q\to qg}\left(\frac{x}{\xi}\right) + K_{qq}\right]\right\} , \tag{140}$$

where K_{qq} denotes finite terms depending on the regularisation scheme. Then the structure function becomes

$$\begin{aligned} F_{2,q}(x,Q^2) = x\sum_i e_{q_i}^2 & \int_x^1 \frac{\mathrm{d}\xi}{\xi} f_i(\xi,\mu_f^2)\, \times \\ & \left\{\delta(1-\frac{x}{\xi}) + \frac{\alpha_s(\mu_r)}{2\pi}\left[P_{q\to qg}\left(\frac{x}{\xi}\right)\ln\frac{Q^2}{\mu_f^2} + \left(C_2^q - K_{qq}\right)\right]\right\} \\ = x\sum_i e_{q_i}^2 & \int_x^1 \frac{\mathrm{d}\xi}{\xi} f_i(\xi,\mu_f^2)\,\hat{F}_{2,i}(\frac{x}{\xi},Q^2,\mu_r,\mu_f) \, . \end{aligned} \tag{141}$$

Defining a convolution in x-space by $f \otimes_x g \equiv \int_x^1 \frac{\mathrm{d}\xi}{\xi} f(\xi)\, g\left(\frac{x}{\xi}\right)$, we see that the structure function is factorised in the form of a convolution,

$$F_{2,q}(x, Q^2) = x \sum_i e_{q_i}^2\, f_i(\mu_f) \otimes_x \hat{F}_{2,i}(\mu_r, t) \;\text{ with }\; t = \ln \frac{Q^2}{\mu_f^2}\;. \tag{142}$$

The long-distance physics is factored into the PDFs which depend on the *factorisation scale* μ_f. The short-distance physics is factored into the hard scattering cross section which depends on both the factorisation and the renormalisation scales. Both scales are arbitrary, unphysical scales. The finite terms depend on the *factorisation scheme*. They are not unique, as finite terms can be shifted between the short- and long-distance parts.

5.4 Parton evolution and the DGLAP equations

With Eq. (142) we again have an equation where an unphysical scale appears on the right-hand side, while the left-hand side is a physical quantity and therefore should not depend on the scale μ_f (when calculated to all orders in perturbation theory). This gives us something akin to a renormalisation group equation, which means that we can calculate how the PDFs evolve as the scale μ_f is changed. As the convolution in Eq. (142) is somewhat inconvenient, we go to Mellin space, where the convolution in the factorisation formula Eq. (142) turns into simple products. The Mellin transform is defined by

$$f(N) \equiv \int_0^1 \mathrm{d}x\, x^{N-1} f(x)\;.$$

The structure function in Mellin space then becomes

$$F_{2,q}(N, Q^2) = x \sum_i e_{q_i}^2\, f_i(N, \mu_f^2)\, \hat{F}_{2,i}(N, \mu_r, t)\;. \tag{143}$$

As a measurable quantity, the structure function must be independent of μ_f, therefore

$$\frac{\mathrm{d}F_{2,q}(N, Q^2)}{\mathrm{d}\mu_f} = 0\;. \tag{144}$$

Note that if F_2 is calculated to order α_s^n, we have $\mu_f^2\, \mathrm{d}F_{2,q}(N, Q^2)/\mathrm{d}\mu_f^2 = \mathcal{O}(\alpha_s^{n+1})$. Therefore, as in the case of the renormalisation scale μ_r, the truncation of the perturbative series introduces a dependence on the unphysical scale in the observable, which gets weaker the more orders we calculate.

For simplicity, let us leave out the sum over i in Eq. (143) and consider only one quark flavour q. We obtain from Eq. (144)

$$\hat{F}_{2,q}(N, t) \frac{\mathrm{d}f_q(N, \mu_f^2)}{\mathrm{d}\mu_f^2} + f_q(N, \mu_f^2) \frac{\mathrm{d}\hat{F}_{2,q}(N, t)}{\mathrm{d}\mu_f^2} = 0\;. \tag{145}$$

Dividing by $f_q\,\hat{F}_{2,q}$ and multiplying by μ_f^2 we obtain

$$\mu_f^2 \frac{\mathrm{d}\ln f_q(N,\mu_f^2)}{\mathrm{d}\mu_f^2} = -\mu_f^2 \frac{\mathrm{d}\ln \hat{F}_{2,q}(N,t)}{\mathrm{d}\mu_f^2} \equiv \gamma_{qq}(N)\,. \tag{146}$$

Using $t = \ln\left(Q^2/\mu_f^2\right)$ this can be written as

$$\frac{\mathrm{d}f_q(N,t)}{\mathrm{d}t} = \gamma_{qq}(N)\, f_q(N,t)\,, \tag{147}$$

where

$$\gamma_{qq}(N) = \int_0^1 \mathrm{d}x\, x^{N-1} P_{qq}(x) = P_{qq}(N)\,. \tag{148}$$

$\gamma_{qq}(N)$ is called the *anomalous dimension* because it measures the deviation of $\hat{F}_{2,q}$ from its naïve scaling dimension. It corresponds to the Mellin transform of the splitting functions.

Very importantly, Eq. (147) implies that the *scale dependence* of the parton densities can be calculated in perturbation theory. The PDFs themselves are non-perturbative, so they have to be extracted from experiment. However, the universality of the PDFs (for each flavour) and the calculable scale dependence means that we can measure the PDFs in one process at a certain scale and then use them in another process at a different scale.

A rigorous treatment based on operator product expansion and the renormalisation group equations extends the above result to all orders in perturbation theory, leading to

$$\boxed{\frac{\partial}{\partial t}\, f_{q_i}(x,t) = \int_x^1 \frac{\mathrm{d}\xi}{\xi}\, P_{q_i/q_j}\left(\frac{x}{\xi}, \alpha_s(t)\right) f_{q_j}(\xi,t)\,.} \tag{149}$$

The splitting functions P_{q_i/q_j} can be calculated as a power series in α_s:

$$P_{q_i/q_j}(x,\alpha_s) = \frac{\alpha_s}{2\pi}\, P_{ij}^{(0)}(x) + \left(\frac{\alpha_s}{2\pi}\right)^2 P_{ij}^{(1)}(x) + \left(\frac{\alpha_s}{2\pi}\right)^3 P_{ij}^{(2)}(x) + \mathcal{O}(\alpha_s^4)\,. \tag{150}$$

Equation (149) holds for distributions which are *non-singlets* under the flavour group: either a single flavour or a combination $q_{\text{ns}} = f_{q_i} - f_{q_j}$ with q_i, q_j being a quark or antiquark of any flavour. More generally, the evolution equation is a $(2n_f + 1)$-dimensional matrix equation in the space of quarks, antiquarks and gluons,

$$\frac{\partial}{\partial t}\begin{pmatrix} f_{q_i}(x,t) \\ f_g(x,t) \end{pmatrix} = \sum_{q_j,\bar{q}_j} \int_x^1 \frac{d\xi}{\xi} \begin{pmatrix} P_{q_i/q_j}(\frac{x}{\xi},\alpha_s(t)) & P_{q_i/g}(\frac{x}{\xi},\alpha_s(t)) \\ P_{g/q_j}(\frac{x}{\xi},\alpha_s(t)) & P_{g/g}(\frac{x}{\xi},\alpha_s(t)) \end{pmatrix} \begin{pmatrix} f_{q_j}(\xi,t) \\ f_g(\xi,t) \end{pmatrix}\,. \tag{151}$$

Equation (151) and (149) are called *DGLAP equations*, named after Dokshitzer [32], Gribov, Lipatov [46] and Altarelli, Parisi [3]. They are among the most important equations in perturbative QCD.

Note that, because of charge conjugation invariance and $SU(N_F)$ flavour symmetry, the splitting functions $P_{q/g}$ and $P_{g/q}$ are independent of the quark flavour and the same for quarks and antiquarks.

Defining the singlet distribution

$$\Sigma(x,t) = \sum_{i=1}^{N_F} [\, f_{q_i}(x,t) + f_{\bar{q}_i}(x,t)\,] \tag{152}$$

and taking into account the considerations above, Eq. (151) simplifies to

$$\frac{\partial}{\partial t}\begin{pmatrix} \Sigma(x,t) \\ g(x,t) \end{pmatrix} = \int_x^1 \frac{d\xi}{\xi} \begin{pmatrix} P_{q/q}(\frac{x}{\xi},\alpha_s(t)) & 2N_F\, P_{q/g}(\frac{x}{\xi},\alpha_s(t)) \\ P_{g/q}(\frac{x}{\xi},\alpha_s(t)) & P_{g/g}(\frac{x}{\xi},\alpha_s(t)) \end{pmatrix} \begin{pmatrix} \Sigma(\xi,t) \\ g(\xi,t) \end{pmatrix} . \tag{153}$$

The leading-order splitting functions including the regulating contributions at $x = 1$ are given by

$$\begin{aligned} P^{(0)}_{q/q}(x) &= C_F\Big\{\frac{1+x^2}{(1-x)_+} + \frac{3}{2}\,\delta(1-x)\Big\} \\ P^{(0)}_{q/g}(x) &= T_R\Big\{x^2 + (1-x)^2\Big\} \quad T_R = \frac{1}{2} \\ P^{(0)}_{g/q}(x) &= C_F\Big\{\frac{1+(1-x)^2}{x}\Big\} \\ P^{(0)}_{g/g}(x) &= 2N_c\Big\{\frac{x}{(1-x)_+} + \frac{1-x}{x} + x(1-x)\Big\} \\ &\quad + \delta(1-x)[\frac{11}{6}N_c - \frac{2}{3}N_F T_R] \;, \end{aligned} \tag{154}$$

where the so-called *plus prescription* is used in Eq. (154), defined by

$$\int_0^1 dx f(x)\left(\frac{1}{1-x}\right)_+ = \int_0^1 dx\,\frac{f(x)-f(1)}{1-x}$$

6 Outlook

Perturbative QCD is a fascinating field with many facets. On the phenomenological side, the impressive experimental tests of the Standard Model achieved to date would have been impossible without the progress in calculating higher-order QCD corrections in the last few decades. Furthermore, important insights have been gained about the infrared structure of QCD at higher orders, the limitations of perturbation theory and the mathematical structure of scattering amplitudes. Equipped with this knowledge, we are well prepared to face the next challenges in collider physics.

Acknowledgements

I would like to thank the organisers of the European Schools of High Energy Physics 2023 for giving me the opportunity to give these lectures and for the excellent organisation. I also would like to thank the discussion leaders for their engagement and the students for their active participation and interesting questions.

References

[1] https://feynrules.irmp.ucl.ac.be/.

[2] A. Accardi et al. Electron Ion Collider: The Next QCD Frontier: Understanding the glue that binds us all. *Eur. Phys. J. A*, 52(9):268, 2016.

[3] Guido Altarelli and G. Parisi. Asymptotic Freedom in Parton Language. *Nucl. Phys.*, B126:298–318, 1977.

[4] Charalampos Anastasiou, Claude Duhr, Falko Dulat, Franz Herzog, and Bernhard Mistlberger. Higgs Boson Gluon-Fusion Production in QCD at Three Loops. *Phys. Rev. Lett.*, 114:212001, 2015.

[5] P. A. Baikov, K. G. Chetyrkin, and J. H. Kühn. Quark Mass and Field Anomalous Dimensions to $\mathcal{O}(\alpha_s^5)$. *JHEP*, 10:076, 2014.

[6] P. A. Baikov, K. G. Chetyrkin, and J. H. Kühn. Five-loop fermion anomalous dimension for a general gauge group from four-loop massless propagators. *JHEP*, 04:119, 2017.

[7] P. A. Baikov, K. G. Chetyrkin, and J. H. Kühn. Five-Loop Running of the QCD coupling constant. *Phys. Rev. Lett.*, 118(8):082002, 2017.

[8] Matteo Becchetti, Roberto Bonciani, Vittorio Del Duca, Valentin Hirschi, Francesco Moriello, and Armin Schweitzer. Next-to-leading order corrections to light-quark mixed QCD-EW contributions to Higgs boson production. *Phys. Rev. D*, 103(5):054037, 2021.

[9] Zvi Bern and David A. Kosower. Color decomposition of one loop amplitudes in gauge theories. *Nucl. Phys. B*, 362:389–448, 1991.

[10] S. Bethke, Z. Kunszt, D. E. Soper, and W. James Stirling. New jet cluster algorithms: Next-to-leading order QCD and hadronization corrections. *Nucl. Phys.*, B370:310–334, 1992. [Erratum: Nucl. Phys.B523,681(1998)].

[11] Georgios Billis, Bahman Dehnadi, Markus A. Ebert, Johannes K. L. Michel, and Frank J. Tackmann. Higgs pT Spectrum and Total Cross Section with Fiducial Cuts at Third Resummed and Fixed Order in QCD. *Phys. Rev. Lett.*, 127(7):072001, 2021.

[12] C. G. Bollini and J. J. Giambiagi. Dimensional Renormalization: The Number of Dimensions as a Regularizing Parameter. *Nuovo Cim.*, B12:20–26, 1972.

[13] Marco Bonetti, Erik Panzer, and Lorenzo Tancredi. Two-loop mixed QCD-EW corrections to $q\overline{q} \to Hg, qg \to Hq$, and $\overline{q}g \to H\overline{q}$. *JHEP*, 06:115, 2022.

[14] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The Anti-k(t) jet clustering algorithm. *JHEP*, 04:063, 2008.

[15] John Campbell and Tobias Neumann. Third order QCD predictions for fiducial W-boson production. *JHEP*, 11:127, 2023.

[16] S. Catani and M. H. Seymour. A General algorithm for calculating jet cross-sections in NLO QCD. *Nucl. Phys.*, B485:291–419, 1997. [Erratum: Nucl. Phys.B510,503(1998)].

[17] S. Catani, L. Trentadue, G. Turnock, and B.R. Webber. Resummation of large logarithms in e+ e- event shape distributions. *Nucl. Phys. B*, 407:3–42, 1993.

[18] Stefano Catani, Simone Devoto, Massimiliano Grazzini, Stefan Kallweit, and Javier Mazzitelli.

Top-quark pair hadroproduction at NNLO: differential predictions with the $\overline{MS}$ mass. *JHEP*, 08(08):027, 2020.

[19] Stefano Catani and Massimiliano Grazzini. The soft gluon current at one loop order. *Nucl. Phys.*, B591:435–454, 2000.

[20] X. Chen, T. Gehrmann, E. W. N. Glover, A. Huss, and J. Mo. NNLO QCD corrections in full colour for jet production observables at the LHC. *JHEP*, 09:025, 2022.

[21] Xuan Chen, Thomas Gehrmann, Nigel Glover, Alexander Huss, Tong-Zhi Yang, and Hua Xing Zhu. Transverse mass distribution and charge asymmetry in W boson production to third order in QCD. *Phys. Lett. B*, 840:137876, 2023.

[22] K. G. Chetyrkin, G. Falcioni, F. Herzog, and J. A. M. Vermaseren. Five-loop renormalisation of QCD in covariant gauges. *JHEP*, 10:179, 2017. [Addendum: JHEP12,006(2017)].

[23] K. G. Chetyrkin and M. Steinhauser. The Relation between the MS-bar and the on-shell quark mass at order alpha(s)**3. *Nucl. Phys. B*, 573:617–651, 2000.

[24] M. Czakon, R. V. Harlander, J. Klappert, and M. Niggetiedt. Exact Top-Quark Mass Dependence in Hadronic Higgs Production. *Phys. Rev. Lett.*, 127(16):162002, 2021. [Erratum: Phys.Rev.Lett. 131, 179901 (2023)].

[25] Michał Czakon, Felix Eschment, Marco Niggetiedt, Rene Poncelet, and Tom Schellenberger. Top-Bottom Interference Contribution to Fully-Inclusive Higgs Production. 12 2023.

[26] Luc Darmé et al. UFO 2.0: the 'Universal Feynman Output' format. *Eur. Phys. J. C*, 83(7):631, 2023.

[27] G. Das, S. Moch, and A. Vogt. Approximate four-loop QCD corrections to the Higgs-boson production cross section. *Phys. Lett. B*, 807:135546, 2020.

[28] Joshua Davies, Matthias Steinhauser, and David Wellmann. Completing the hadronic Higgs boson decay at order α_s^4. *Nucl. Phys.*, B920:20–31, 2017.

[29] Vittorio Del Duca, Lance J. Dixon, and Fabio Maltoni. New color decompositions for gauge amplitudes at tree and loop level. *Nucl. Phys. B*, 571:51–70, 2000.

[30] G. Dissertori, I. G. Knowles, and M. Schmelling. *Quantum Chromodynamics: High energy experiments and theory*. International Series of Monographs on Physics No. 115, Oxford University Press, 2005.

[31] Lance J. Dixon. A brief introduction to modern amplitude methods. In *Theoretical Advanced Study Institute in Elementary Particle Physics: Particle Physics: The Higgs Boson and Beyond*, pages 31–67, 2014.

[32] Yuri L. Dokshitzer. Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics. *Sov. Phys. JETP*, 46:641–653, 1977. [Zh. Eksp. Teor. Fiz.73,1216(1977)].

[33] Yuri L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber. Better jet clustering algorithms. *JHEP*, 08:001, 1997.

[34] Claude Duhr, Falko Dulat, and Bernhard Mistlberger. Charged current Drell-Yan production at N^{3}LO. *JHEP*, 11:143, 2020.

[35] Falko Dulat, Achilleas Lazopoulos, and Bernhard Mistlberger. iHixs 2 — Inclusive Higgs cross sections. *Comput. Phys. Commun.*, 233:243–260, 2018.

[36] Matteo Fael, Fabian Lange, Kay Schönwald, and Matthias Steinhauser. A semi-analytic method to compute Feynman integrals applied to four-loop corrections to the $\overline{\mathrm{MS}}$-pole quark mass relation. *JHEP*, 09:152, 2021.

[37] J. Fleischer, F. Jegerlehner, O. V. Tarasov, and O. L. Veretin. Two loop QCD corrections of the massive fermion propagator. *Nucl. Phys. B*, 539:671–690, 1999. [Erratum: Nucl.Phys.B 571, 511–512 (2000)].

[38] H. Fritzsch, Murray Gell-Mann, and H. Leutwyler. Advantages of the Color Octet Gluon Picture. *Phys. Lett.*, 47B:365–368, 1973.

[39] Einan Gardi and Georges Grunberg. Power corrections in the single dressed gluon approximation: The Average thrust as a case study. *JHEP*, 11:016, 1999.

[40] A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, and G. Heinrich. NNLO corrections to event shapes in e+ e- annihilation. *JHEP*, 12:094, 2007.

[41] A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, and G. Heinrich. Jet rates in electron-positron annihilation at O(alpha(s)**3) in QCD. *Phys. Rev. Lett.*, 100:172001, 2008.

[42] A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, A. Huss, and J. Pires. Triple Differential Dijet Cross Section at the LHC. *Phys. Rev. Lett.*, 123(10):102001, 2019.

[43] Murray Gell-Mann. A Schematic Model of Baryons and Mesons. *Phys. Lett.*, 8:214–215, 1964.

[44] N. Gray, David J. Broadhurst, W. Grafe, and K. Schilcher. Three Loop Relation of Quark (Modified) Ms and Pole Masses. *Z. Phys. C*, 48:673–680, 1990.

[45] Massimiliano Grazzini, S. Kallweit, Jonas M. Lindert, Stefano Pozzorini, and Marius Wiesemann. NNLO QCD + NLO EW with Matrix+OpenLoops: precise predictions for vector-boson pair production. *JHEP*, 02:087, 2020.

[46] V. N. Gribov and L. N. Lipatov. Deep inelastic e p scattering in perturbation theory. *Sov. J. Nucl. Phys.*, 15:438–450, 1972. [Yad. Fiz.15,781(1972)].

[47] David J. Gross and Frank Wilczek. Ultraviolet Behavior of Nonabelian Gauge Theories. *Phys. Rev. Lett.*, 30:1343–1346, 1973. [,271(1973)].

[48] Franz Gross et al. 50 Years of Quantum Chromodynamics. *Eur. Phys. J. C*, 83:1125, 2023.

[49] Andrey G. Grozin, Peter Marquard, Alexander V. Smirnov, Vladimir A. Smirnov, and Matthias Steinhauser. Matching the heavy-quark fields in QCD and HQET at four loops. *Phys. Rev. D*, 102(5):054008, 2020.

[50] Robert V. Harlander and Matthias Steinhauser. rhad: A Program for the evaluation of the hadronic R ratio in the perturbative regime of QCD. *Comput. Phys. Commun.*, 153:244–274, 2003.

[51] A. Heister et al. Studies of QCD at e+ e- centre-of-mass energies between 91-GeV and 209-GeV. *Eur. Phys. J. C*, 35:457–486, 2004.

[52] F. Herzog, B. Ruijl, T. Ueda, J. A. M. Vermaseren, and A. Vogt. The five-loop beta function of Yang-Mills theory with fermions. *JHEP*, 02:090, 2017.

[53] Gregor Kälin. Cyclic Mario worlds — color-decomposition for one-loop QCD. *JHEP*, 04:141,

2018.

[54] Vardan Khachatryan et al. Measurement and QCD analysis of double-differential inclusive jet cross sections in pp collisions at $\sqrt{s} = 8$ TeV and cross section ratios to 2.76 and 7 TeV. *JHEP*, 03:156, 2017.

[55] T. Kinoshita. Mass singularities of Feynman amplitudes. *J. Math. Phys.*, 3:650–677, 1962.

[56] Stefan Kluth. Tests of Quantum Chromo Dynamics at e+ e- Colliders. *Rept. Prog. Phys.*, 69:1771–1846, 2006.

[57] E. Laenen. QCD. In *Proceedings, 2014 European School of High-Energy Physics (ESHEP 2014): Garderen, The Netherlands, June 18 - July 01 2014*, pages 1–58, 2016.

[58] T. D. Lee and M. Nauenberg. Degenerate Systems and Mass Singularities. *Phys. Rev.*, 133:B1549–B1562, 1964. [,25(1964)].

[59] K. Long. QCD at high-energy (experiments). 2002. [Nucl. Phys. Proc. Suppl.117,242(2003)].

[60] Thomas Luthe, Andreas Maier, Peter Marquard, and York Schröder. Five-loop quark mass and field anomalous dimensions for a general gauge group. *JHEP*, 01:081, 2017.

[61] Thomas Luthe, Andreas Maier, Peter Marquard, and York Schröder. The five-loop Beta function for a general gauge group and anomalous dimensions beyond Feynman gauge. *JHEP*, 10:166, 2017.

[62] F. Maltoni, K. Paul, T. Stelzer, and S. Willenbrock. Color Flow Decomposition of QCD Amplitudes. *Phys. Rev. D*, 67:014026, 2003.

[63] Peter Marquard, Alexander V. Smirnov, Vladimir A. Smirnov, Matthias Steinhauser, and David Wellmann. $\overline{\mathrm{MS}}$-on-shell quark mass relation up to four loops in QCD and a general $\mathrm{SU}(N)$ gauge group. *Phys. Rev. D*, 94(7):074025, 2016.

[64] Simone Marzani, Gregory Soyez, and Michael Spannowsky. *Looking inside jets: an introduction to jet substructure and boosted-object phenomenology*, volume 958. Springer, 2019.

[65] Tom Melia. Proof of a new colour decomposition for QCD amplitudes. *JHEP*, 12:107, 2015.

[66] Kirill Melnikov and Timo van Ritbergen. The Three loop relation between the MS-bar and the pole quark masses. *Phys. Lett. B*, 482:99–108, 2000.

[67] Bernhard Mistlberger. Higgs boson production at hadron colliders at N^{3}LO in QCD. *JHEP*, 05:028, 2018.

[68] Alexander Ochirov and Ben Page. Full Colour for Loop Amplitudes in Yang-Mills Theory. *JHEP*, 02:100, 2017.

[69] H. David Politzer. Reliable Perturbative Results for Strong Interactions? *Phys. Rev. Lett.*, 30:1346–1349, 1973. [,274(1973)].

[70] Ted C. Rogers. An overview of transverse-momentum–dependent factorization and evolution. *Eur. Phys. J. A*, 52(6):153, 2016.

[71] Gavin P. Salam. Towards Jetography. *Eur. Phys. J. C*, 67:637–686, 2010.

[72] Peter Skands. Introduction to QCD. In *Theoretical Advanced Study Institute in Elementary Particle Physics: Searching for New Physics at Small and Large Scales*, pages 63–124, 2017.

[73] George F. Sterman and Steven Weinberg. Jets from Quantum Chromodynamics. *Phys. Rev. Lett.*, 39:1436, 1977.

[74] Gerard 't Hooft and M. J. G. Veltman. Regularization and Renormalization of Gauge Fields. *Nucl. Phys.*, B44:189–213, 1972.

[75] Z. Trocsanyi. QCD for collider experiments. In *Proceedings, 2013 European School of High-Energy Physics (ESHEP 2013): Paradfurdo, Hungary, June 5-18, 2013*, pages 65–116, 2015.

[76] T. van Ritbergen, J. A. M. Vermaseren, and S. A. Larin. The Four loop beta function in quantum chromodynamics. *Phys. Lett.*, B400:379–384, 1997.

[77] R. L. Workman et al. Review of Particle Physics. *PTEP*, 2022:083C01, 2022.

[78] G Zweig. An SU_3 model for strong interaction symmetry and its breaking; Version 1. Technical Report CERN-TH-401, CERN, Geneva, Jan 1964.

Cosmology

Mikhail Shaposhnikov[a]

[a]EPFL, Lausanne, Switzerland

This series of lectures covers the basics of cosmology from a particle physics point of view. The following topics will be partially covered: expanding Universe, cosmological parameters, generic approach to physical processes in the early Universe, cosmic microwave background radiation, nucleosynthesis, baryogenesis, dark matter, cosmological phase transitions and inflation.

This chapter should be cited as: Cosmology, Mikhail Shaposhnikov, DOI: 10.23730/CYRSP-2025-009.65,
in: Proceedings of the 2023 European School of High-Energy Physics,
CERN Yellow Reports: School Proceedings, CERN-2025-009, DOI: 10.23730/CYRSP-2025-009, p.65.

1 Introduction

At first sight, cosmology and particle physics seem to be unrelated branches of physics. Particle physics aims to describe elementary particles and their fundamental interactions at small scales, say, $l < 10^{-14}$ cm. On the contrary, the goal of cosmology is to describe the structure of the Universe at enormous length scales, $l > 10\ \mathrm{kpc} \simeq 10^{22}$ cm. Can we learn anything from cosmology for particle physics? Can we learn anything from particle physics for cosmology?

The evolution of the Universe provides the bridge between particle physics and cosmology. The Universe is expanding, meaning that it was very dense in the past. The physics of the early Universe depends crucially on the properties of elementary particles and the interactions between them. The observations of the present Universe can give us information about the early Universe, and, therefore, about particle physics. The composition of the Universe provides us with the strongest indication that the Standard Model (SM) of elementary particles is not complete, as it cannot explain why we have more matter than antimatter and what is Dark Matter (DM). Cosmology may supply us with constraints on particle physics theories, sometimes superior to those coming from terrestrial experiments. The list of cosmological insights into high energy physics also includes bounds on neutrino masses and the number of light particle species, and constraints on the properties of hypothetical particles and their interactions.

On the other hand, progress in particle physics has led to many advances in cosmology. The non-conservation of baryon number already present in the Standard Model unified with the new sources of CP violation existing in many of its extensions, has led to a qualitative understanding of the absence of antimatter in the Universe; new stable particles, predicted by different Beyond the SM (BSM) models, may play the role of dark matter in the Universe; consideration of phase transitions in particle-physics models has led to the suggestion of a new paradigm in cosmology - inflation. So a “simple” thing as the dynamics of a free quantum scalar field in the expanding Universe proposes a solution to several outstanding problems in cosmology, such as flatness, horizon, homogeneity and structure formation.

The importance of cosmology for particle physics has risen immensely during the last few years. Indeed, the experiments in neutrino physics and robust conclusions of modern observational cosmology call for extensions of the Standard Model of particle physics. However, the situation is very different from what we had in the years preceding the discovery of the Higgs boson. The consistency of the SM, together with experimental results that existed at the time, allowed us to firmly conclude that either the Higgs boson had to be discovered at the LHC, or new physics beyond the SM must show up. Currently, we know for sure that the SM is incomplete and has to be extended. However, we do not have a firm prediction from particle theory of where to search for new particles and what their masses, spin, and interaction type could be.

The famous cosmologist, Zeldovich, used to say: “The Universe is a poor men accelerator - it can produce very heavy particles and very weakly interacting particles. Unfortunately, the experiment happened just once.” In the absence of hints from particle physics experiments for new physics, cosmology may help to identify the future directions.

The plan of the lectures is as follows. First, we are going to remind the necessary elements of General Relativity. Then we will go through the basic facts about the Universe: its large-scale isotropy and homogeneity, its Hubble expansion, and learn how to describe the Universe's evolution with the Friedman equations. We will consider the content of the Universe and see that its composition presents a puzzle which cannot be explained by the Standard Model of particle interactions. Then we will turn to the foundations of the Big Bang Theory and set up a semi-quantitative way to analyse the processes in the early Universe, based on the notions of thermal equilibrium and non-equilibrium, on freeze-in and freeze-out of particle reactions. After that, we will use these tools to discuss the origins of CMB (Cosmic Microwave Background), photon and neutrino decoupling in the early Universe, nucleosynthesis, and baryogenesis, leading to a matter-antimatter asymmetric Universe. Next, we will review the evidence for Dark Matter and Dark energy and describe one of the possible Dark Matter candidates. The last topic is inflation: we shall discuss the problems of the standard cosmological model and how they can be solved by cosmological inflation.

This writeup is **not self-contained**, it should be used together with the slides shown at the School. In particular, most of the figures are not reproduced here.

Unless otherwise specified, we will use the natural system of units, in which $\hbar = c = 1$ and energy is measured in GeV.

There are several excellent textbooks on cosmology (e.g. [1–6]) that a reader can consult for a thorough study of the subject. Another source of information is https://pdg.lbl.gov, where updated reviews on Astrophysics and Cosmology can be found. No references to the original works will be given here, see the resources listed above and reviews cited in the text.

2 Some elements of General Relativity

In three-dimensional Euclidean space, the infinitesimal distance dl in Cartesian coordinates x, y and z is written as

$$dl^2 = dx^2 + dy^2 + dz^2 \,. \tag{1}$$

This expression is invariant under symmetries of the Euclidean space, namely rotations and translations. The object which is used for the construction of a relativistic theory is the interval ds,

$$ds^2 = dt^2 - dl^2 = \eta_{\mu\nu} dx^\mu dx^\nu \,, \tag{2}$$

where $t = x^0$ is the time coordinate, and the Minkowski metric corresponding to the flat space-time is given by the diagonal matrix $\eta_{\mu\nu} = diag(1, -1, -1, -1)$. In addition to translations and rotations, the interval is invariant under Lorentz transformations, corresponding to transition between different inertial frames.

The generalisation to curved space-time is associated with the interval expressed through the general metric $g_{\mu\nu}$ which is some function of arbitrary coordinates x^μ,

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \,, \tag{3}$$

The metric $g_{\mu\nu}$ is now a dynamical variable, describing the non-trivial geometry of the space-time which is determined by matter and the gravitational field itself. Einstein's theory of gravity is constructed in such a way that its field-theory action is invariant under general coordinate transformation,

$$x^\mu \to x'^\mu = f^\mu(x)\,, \tag{4}$$

where $f^\mu(x)$ are arbitrary functions of coordinates. This ensures, in particular, the validity of the equivalence principle (inertial mass is the same as the gravitational mass).

The action of the 'theory of everything" can be symbolically written as

$$S = S_{\text{gravity}} + S_{\text{matter}}\,, \tag{5}$$

where S_{matter} includes the Lagrangian of the Standard Model and yet other pieces remain to be determined, whereas the lowest order gravity action is that of Einstein-Hilbert,

$$S_{\text{gravity}} = -\frac{1}{8\pi G}\int d^4x\sqrt{-g}\left(\frac{R}{2}+\lambda\right)\,, \tag{6}$$

where g is the determinant of the metric, G is the Newton constant of gravity, λ is the cosmological constant, and R is the scalar curvature constructed out of the Riemann curvature tensor as $R = g^{\nu\sigma}R^\mu_{\nu\mu\sigma}$. The equations of motion, as usual, are derived by the variation of the action with respect to the metric. These are the famous Einstein equations

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R - \lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}\,, \tag{7}$$

where $R_{\mu\nu}$ is the Ricci tensor, and $T_{\mu\nu}$ is the matter stress-energy tensor. These equations are capable of describing all gravitational phenomena observed till now. We will use them for the analysis of the expanding Universe. It is customary to absorb the cosmological constant into the matter tensor as

$$\tilde{T}_{\mu\nu} = T_{\mu\nu} + \frac{\lambda}{8\pi G}g_{\mu\nu}\,. \tag{8}$$

This operation makes the Einstein equations look like

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = 8\pi G T_{\mu\nu}\,, \tag{9}$$

where we removed the tilde to simplify the notations.

3 Structure of the Universe at large distances and the FRW metric

To give a sense of scales for various objects studied in cosmology and astrophysics, we list here the sizes of some of the most recognisable structures in the Universe:

- Earth radius $\approx 6.4 \times 10^8$ cm.
- Solar radius $\approx 7.0 \times 10^{10}$ cm.
- Earth-Sun distance $\equiv$ 1AU $\approx 1.5 \times 10^{13}$ cm.

To go beyond this point, it is useful to adopt a new unit of distance called parsec

$$1 \text{ pc} = \frac{1 \text{ AU}}{1 \text{ arcsec}} = 3.26 \text{ lightyear} = 3 \times 10^{18} \text{ cm}. \tag{10}$$

– A galaxy can be thought of as a disk of radius ~ 30 kpc and thickness ~ 1 kpc.

– The galaxies in our Universe are clumped together in clusters of typical size ~ 10 Mpc. The Virgo galaxy cluster, for example, consists of $\sim 10^3 - 10^4$ galaxies.

– Finally, the size of our observable Universe is $\sim 5 \times 10^3$ Mpc.

The most important observational fact about the Universe is that it is isotropic and homogeneous on large scales, i.e. much larger than the galaxy cluster scale. The isotropy of the Universe can be verified by counting and comparing the number of galaxies in different directions or by observing the cosmic microwave background. To prove the homogeneity of the Universe one should reconstruct its three-dimensional picture by measuring distances between galaxies.

The mathematical description of the homogeneous and isotropic Universe is based on the metric which is known as the Friedmann-Lemaitre-Robertson-Walker metric or simply the FLRW metric or even FRW metric:

$$ds^2 = dt^2 - a^2(t)d\bar{l}^2\,, \;\; d\bar{l}^2 = \left(\frac{d\bar{r}^2}{1-\kappa\bar{r}^2} + \bar{r}^2 d\Omega^2\right)\,, \tag{11}$$

where $\bar{r}$ is the radial coordinate, $d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$ is the solid angle, and the parameter κ describes the global characteristic of the Universe:

$$\kappa = \begin{cases} +1, & \text{closed, positive curvature} \\ \;\;0, & \text{flat, zero curvature} \\ -1, & \text{open, negative curvature}\,. \end{cases} \tag{12}$$

The function $a(t)$ is called the scale factor. Increasing with time $a(t)$ leads to an expanding Universe, whereas if $a(t)$ decreases the Universe would be collapsing.

The structure of the FRW metric allows us to establish the Hubble law for the expanding Universe. If the galaxies are not too far from each other, the physical distance l between them can be written as

$$l = a(t)d\bar{l}\,, \tag{13}$$

leading to

$$\dot{l} = \dot{a}(t)d\bar{l}\,, \tag{14}$$

where the dot is the time derivative (note that the coordinate distance $d\bar{l}$ is time-independent). Excluding $d\bar{l}$, we arrive to the Hubble law

$$\dot{l} = Hl\,, \;\; \text{or} \;\; \vec{v} = H\vec{r}\,, \tag{15}$$

saying that the running out velocity $\vec{v}$ is proportional to the distance $\vec{r}$. Here $H = \dot{a}/a$ is the Hubble constant.

The running-out velocity and the distance can be traded off to the more convenient quantities, which can be experimentally measured. One of them is the "redshift" z and another is the "luminosity distance" d.

There are several types of objects in the Universe that can be considered as "standard candles". This means that we know their total luminosity L and the spectrum of emitted light *at the position they find themselves*. Examples of such objects include supernovae of the type Ia, first-ranked E galaxies in nearby groups and clusters, first-ranked cluster galaxies in rich clusters, etc. With this knowledge, we can find for these objects the redshifts z, defined as:

$$z = \frac{\lambda_{\rm rec} - \lambda_{\rm emit}}{\lambda_{\rm emit}} \, , \tag{16}$$

where $\lambda_{\rm emit}$ and $\lambda_{\rm rec}$ are the wavelengths of the emitted and received light respectively.

Simultaneously, one can find the luminosity distance r to the corresponding object, by measuring the energy flux f (apparent brightness) from the star or galaxy, through the equation

$$f = \frac{L}{4\pi d^2} \, , \tag{17}$$

For distances much smaller than the size of the visible Universe the luminosity distance is the same as the physical distance, whereas the redshift can be associated with the Doppler effect since the relation between the frequency of emitted and received light is given by

$$z = \sqrt{\frac{1+v}{1-v}} - 1 \simeq v \, , \tag{18}$$

for $v \ll 1$, where v is the relative velocity of the emitter to the receiver. This means that the relation between z and d, in this case, reads $z = Hd$. It has a universal character and does not depend on the matter content of the Universe, on the type of the object, on the frequency of the emitted light or on the direction in the sky. It has been verified in numerous observations, confirming the Hubble law.

If the luminosity distance is large, the Hubble law starts to depend on the composition of the Universe. The speed of light is finite, observation of the sources at large distances means that we observe them in the past. Another way to write the redshift is

$$z = \frac{a_{\rm now}}{a_{\rm emit}} - 1 \, , \tag{19}$$

where $a_{\rm now}$ and $a_{\rm emit}$ are the scale factors of the Universe at present and at the time the light was emitted. This follows from the fact that the frequency ν of light changes in an expanding Universe in such a way that $\nu a =$ const, which is easy to understand because the product νa just shows the number of wavelengths in a box of the size a and this number does not change if the size of the box changes.

A somewhat more involved consideration allows one to relate the luminosity distance d to the coordinate distance r in the FRW metric. It reads

$$d = \bar{r}\frac{a_{\text{now}}^2}{a_{\text{emit}}} \,. \tag{20}$$

The generalised Hubble law, relating d and z can be found from Equations (19,20), it depends on the way the Universe evolved from the time the light was emitted and absorbed. This, in turn, is a function of the content of the Universe via Einstein's equations (9), which for the homogeneous and isotropic Universe are simplified to the Friedmann equations, discussed below.

4 The Friedmann equations

Since the Universe is homogeneous and isotropic, the stress-energy tensor in (9) must obey the same properties. It is a good approximation to assume that the Universe is a perfect fluid. Its energy-momentum tensor is then

$$T_{\mu\nu} = (\rho + p)u_\mu u_\nu - p g_{\mu\nu} \,, \tag{21}$$

with ρ, p, and u_μ being the energy density, pressure, and 4-velocity of the fluid respectively. The Universe is not boost invariant and has a preferred cosmic rest frame, namely the rest frame of the fluid, in which $u_\mu = (1, 0, 0, 0)$. In this coordinate system, the non-zero components of the stress-energy tensor are

$$T_{00} = \rho \,, \tag{22}$$

$$T_{ij} = -p g_{ij} \,. \tag{23}$$

Using this form of the stress tensor we arrive at two equations, being the 00^{th} and ij^{th} components of Equation (9):

$$\frac{\dot{a}^2}{a^2} + \frac{\kappa}{a^2} - \frac{\Lambda}{3} = \frac{8\pi G}{3}\rho \,, \tag{24}$$

$$2\frac{\ddot{a}}{a} + \frac{\dot{a}^2}{a^2} + \frac{K}{a^2} - \Lambda = -8\pi G p \,. \tag{25}$$

All the other components of the Einstein equation are identically zero. Note that as a consequence of the homogeneity and isotropy of the Universe, the ten Einstein equations have been reduced to only two. Equations (24) and (25) are known as the *Friedmann equations*.

The two Friedmann equations have three unknowns: ρ, p, and a, i.e. they are under-determined. For them to be solvable, one more equation is needed. The extra equation complementing them is usually taken to be the equation of the state of the Universe: an equation of the form $\rho = wp$, where w depends on the properties of the energy component in question. Below are three types of energy components that are of special importance in cosmology together with their equations of states:

- Non-relativistic matter/dust: $p = 0$.
- Relativistic matter/radiation: $p = \rho/3$.
- Cosmological constant/dark energy: $p = -\rho$ (following from $T_{\mu\nu} = \frac{\lambda}{8\pi G} g_{\mu\nu}$).

If several species are present, the equation state is given by the sum of each component

$$p = \sum_i w_i \rho_i \, . \tag{26}$$

Let us take $\kappa = 0$ for simplicity and consider several important cases.

(i) The radiation-dominated Universe, $p = \epsilon/3$. Here

$$\epsilon = \frac{3}{32\pi G}\frac{1}{t^2} \, , \; a = a_0 \left(\frac{t}{t_0}\right)^{\frac{1}{2}} \, , \; H = \frac{1}{2t} \, . \tag{27}$$

(ii) The matter-dominated Universe, $p = 0$. Here

$$\epsilon = \frac{1}{6\pi G}\frac{1}{t^2} \, , \; a = a_0 \left(\frac{t}{t_0}\right)^{\frac{2}{3}} \, , \; H = \frac{2}{3t} \, . \tag{28}$$

(ii) The vacuum-energy-dominated Universe, $\epsilon = -p, \; \epsilon > 0$. Here

$$\epsilon = \text{ const} \, , \; a = a_0 \exp{(Ht)} \, , \; H = \text{ const} = \sqrt{\frac{8\pi G\epsilon}{3}} \, . \tag{29}$$

The last equation may look counter-intuitive since, despite the expansion of the Universe, the energy density does not change. This is related to the fact that the vacuum pressure is negative and it performs negative work which keeps the energy density exactly constant.

A more general case is a mixture of radiation, non-relativistic matter and the vacuum-energy density. Let us introduce different densities, specific for each type of matter,

$$\Omega_M = \frac{\epsilon_M}{\rho_c}, \;\; \Omega_r = \frac{\epsilon_r}{\rho_c}, \;\; \Omega_\Lambda = \frac{\epsilon_\Lambda}{\rho_c} \, , \tag{30}$$

where the indices $M, \; r$ and Λ refer to the contributions of matter, radiation and vacuum energy respectively. The parameter ρ_c called the critical density, is given by

$$\rho_c = \frac{3H_0^2}{8\pi G} \, . \tag{31}$$

where index 0 refers to the present time. The Hubble constant H_0 is usually parameterized as $H_0 = 100 \; h \, \frac{\text{km}}{\text{s}\cdot\text{Mpc}}$ where h is taken from observations, $h \approx 0.67$. Numerically, the critical density is $\rho_c = 1.88h^2 \cdot 10^{-29}$ g/cm^{-3}.

As the Universe expands, different components of the energy scale in the following way:

$$\epsilon_M \sim a^{-3} \, , \;\; \epsilon_r \sim a^{-4} \, , \;\; \epsilon_\Lambda \sim \text{ const} \, , \tag{32}$$

which follows from Equations (27,28,29) and is easy to understand. The equation for matter tells us that the total energy of non-relativistic matter is conserved, the equation for radiation shows that the total number of photons and other light particles is conserved, while their energy is red-shifted. Thus, Equation (24) can be written as

$$H^2 = H_0^2 \left(\Omega_r \frac{a_0^4}{a^4} + \Omega_M \frac{a_0^3}{a^3} + \Omega_\Lambda + \Omega_k \frac{a_0^2}{a^2} \right) , \tag{33}$$

where the curvature contribution, $\Omega_k \equiv \frac{k}{a_0^2 H_0^2}$ has been introduced for uniformity of notation. As before, the index 0 refers to the present moment of the Universe expansion and we have

$$\Omega_r + \Omega_M + \Omega_\Lambda + \Omega_k = 1 . \tag{34}$$

The dominant component of energy density in the early Universe is that related to radiation. Later on, matter dominates. The curvature contribution, potentially important for the evolution of the Universe at a later time, happens to be numerically unimportant, and the Λ term dominates. The schematic dependence of the scale factor on time is represented in Figure 1. The moment when the matter energy density starts to dominate the radiation, $\Omega_M = \Omega_r$ is important for structure formation. This happens at redshift z_{eq} approximately equal to $z_{\text{eq}} = 3.1 \cdot 10^4 \Omega_M h^2 \simeq 3500$ and corresponds to the age of the Universe $t_{\text{eq}} = 5.2 \cdot 10^4$ years.

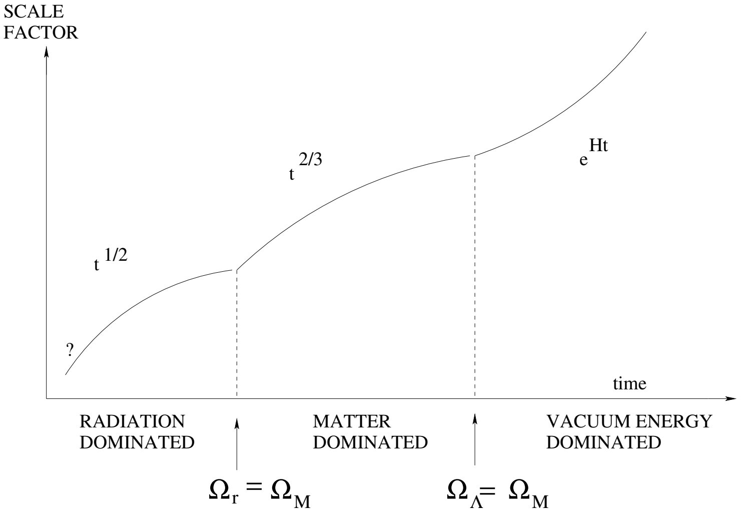


Fig. 1: Dependence of the scale factor on time.

The analysis of the Hubble law at large redshifts based on supernovae, the CMB, the baryon acoustic oscillations, and the dynamics of clusters allows us to pin down the cosmological abundances, giving

$$\Omega_\Lambda \approx 0.68 , \;\; \Omega_M \approx 0.32 , \;\; \Omega_{\text{DM}} \approx 0.27 , \;\; \Omega_{\text{baryon}} \approx 0.05 , \tag{35}$$

where we introduced Ω_{DM} and Ω_{baryon} for the DM and baryon abundances respectively (they are discussed in Sections 11,12).

With this set of parameters, the Universe accelerates and will expand forever. The total energy density in the Universe is close to the critical one. This could be considered as an indication of the validity of the inflationary-Universe scenario, discussed below. The age of the Universe, for $\Omega_{\text{tot}} = 1$ is $\simeq 13.8$ Gyr, and the value of the Hubble constant is $H_0 \simeq 67$ km s^{-1}Mpc^{-1}.

All these numbers (35) represent a puzzle. If the Standard Model was right, one would expect to have $\Omega_{\rm DM} \ll 1$ and $\Omega_{\rm baryon} \ll 1$ (see Sections 11,12)).

5 Big Bang theory

The common feature of the matter and radiation-dominated regimes of the Universe expansion is the mathematical singularity of the solutions (27,28) at $t \to 0$: the scale factor of the Universe goes to zero, $a \to 0$ and the energy density goes to infinity, $\rho \to \infty$. Physically, all the formulas in this limit do not make any sense: the analysis carried out so far was classical and is not valid when the quantum gravity effects become essential, namely at $\rho \sim M_{Pl}^4$. Here M_{Pl} is the Planck mass, related to Newton's gravitational constant as $M_{Pl} = G^{-\frac{1}{2}} \simeq 1.2 \cdot 10^{19}$ GeV. We do not know what happened then, but the solutions show that the energy density of the Universe in the past was much higher than today. The Universe looked like a plasma of different elementary particles which were close to each other. Hence, reactions between particles were rapid enough, which means that the system was driven to a state of thermal equilibrium and had maximal entropy. This is the essence of the Gamow's Big Bang Theory.

This picture leads immediately to several predictions. One of them is the existence of the CMB - cosmic microwave background radiation which has the Planck thermal spectrum. The other is Big Bang nucleosynthesis, leading to the prediction of abundances of light elements, such as ^4He. The CMB has a thermal equilibrium Planck spectrum with temperature $T = 2.73$ K,

$$dI_\nu \sim \frac{\nu^3 d\nu}{\exp\left(\frac{\nu}{T}\right) - 1}\,, \tag{36}$$

where I_ν is the energy density. The CMB we observe today is simply the ensemble of the relic photons, red-shifted to the present time.

The CMB is highly isotropic. Its dipole anisotropy tells us that the Earth is moving through the CMB with velocity $\simeq 370$ km/s; its multipole anisotropies are at the level of 10^{-5} signal the presence of primordial density perturbations, important for the structure formation (see Section 14.4).

6 Thermal equilibrium

The main assumption of the Big Bang theory about thermal equilibrium is very powerful. Indeed, to describe a generic state of a homogeneous system one would need to know the distribution functions of every particle species, giving the number of particles with a specific momentum. If the system is in thermal equilibrium, it is sufficient to tell what are its temperature T and chemical potentials, corresponding to conserved quantum numbers.

The equilibrium non-interacting particle number distributions are

$$n(p) = \frac{1}{e^{\frac{E-\mu}{T}} \pm 1}\,, \tag{37}$$

where the plus sign refers to fermions and the minus sign to bosons, E is the energy of the particle and μ is the chemical potential. The values of the chemical potentials, at least in the radiation-dominated epoch, are rather small and may be omitted for most purposes.

The number density N and energy density ρ of free particles in thermal equilibrium are given by

$$N = \frac{g}{(2\pi)^3} \int d^3\mathbf{p} n(p)\,, \tag{38}$$

$$\rho = \frac{g}{(2\pi)^3} \int d^3\mathbf{p} E(p) n(p)\,, \tag{39}$$

where g is the number of spin states of a particle, $E(p)$ is the energy of a particle with momentum p. In the relativistic limit $T \gg m$, the integrals in (38) and (39) can be calculated explicitly. The results are

$$\rho = \begin{cases} \frac{\pi^2}{30} g T^4 & \text{boson} \\ \frac{7}{8}\frac{\pi^2}{30} g T^4 & \text{fermion} \end{cases} \tag{40}$$

$$N = \begin{cases} \frac{\zeta(3)}{\pi^2} g T^3 & \text{boson} \\ \frac{3}{4}\frac{\zeta(3)}{\pi^2} g T^3 & \text{fermion}\,, \end{cases} \tag{41}$$

where $\zeta(x)$ is the Riemann zeta function, $\zeta(3) \approx 1.2$. If both bosons and fermions are present, the total energy density can be written compactly as

$$\rho = \frac{\pi^2}{30} g_* T^4\,, \tag{42}$$

where

$$g_* \equiv \sum_{\text{bosons}} + \frac{7}{8} \sum_{\text{fermions}}\,, \tag{43}$$

is the effective number of relativistic degrees of freedom. The sums above are taken over species that are relativistic at temperature T. For example, for the Standard Model $g^* = 106.75$, if the temperature is well above the Fermi scale. This equation allows one to write a relation between the temperature and the expansion time, combining Equations (27) and (40):

$$t = \frac{1}{2H} = 0.301 \frac{M_{Pl}}{\sqrt{g_*} T^2} \equiv \frac{M_0}{T^2}\,. \tag{44}$$

To appreciate the orders of magnitude, here is an equation to remember:

$$t[\mathrm{s}] = 1/\mathrm{T}[\mathrm{MeV}]^2\,. \tag{45}$$

In the non-relativistic limit, $T \ll m$, the particle-number densities are

$$n = g \left(\frac{mT}{2\pi}\right)^{\frac{3}{2}} e^{-\frac{m-\mu}{T}}\,. \tag{46}$$

This leads to the energy density $\epsilon = mn$ and to the pressure $p \sim nT \ll \epsilon$. Yet another important thermodynamic quantity is the entropy density, given by

$$s = \frac{2\pi^2}{45} g_* T^3 \tag{47}$$

in the relativistic limit.

7 Freeze in and freeze out

The evolution of the number density N of particles in an expanding Universe is driven by two effects: the diluting effect of the expansion, whose strength is characterized by the Hubble parameter H, and the thermalising effect of particle collisions, whose strength is characterized by the reaction rate Γ

$$\Gamma = \frac{1}{\langle \sigma N v \rangle}\,, \tag{48}$$

where σ is a cross-section of the reaction, N is the relevant particle density, and v is the relative velocity of colliding particles.

The temperature T_* at which

$$\Gamma(T_*) = H(T_*) \tag{49}$$

marks the transition between two qualitatively different behaviours of the number density N. In one extreme $\Gamma(T) \gg H(T)$, the expansion of the Universe is negligible and thermal equilibrium can be achieved without hurdle. In the other extreme $\Gamma(T) \ll H(T)$, interactions happen so rarely that the number of particles in a comoving volume essentially freezes and hence the number density simply scales as $N \propto 1/a^3$. These behaviours are summarized below

$$N \approx \begin{cases} N_{\text{eq}}(T), & \Gamma(T) \gg H(T) \\ N_{\text{eq}}(T_*)\left(\dfrac{a(T_*)}{a(T)}\right)^3, & \Gamma(T) \ll H(T)\,. \end{cases} \tag{50}$$

In the expression for $\Gamma(T) \ll H(T)$, we have picked the point $T = T_*$ as the scaling reference for N. Since at that point $\Gamma(T_*) = H(T_*)$, and not $\Gamma(T_*) \ll H(T_*)$, the expression is only correct approximately, up to an O(1) factor. The temperature T at which $H = \Gamma$ is called *freeze-in* temperature or *freeze-out* temperature, depending on how the ratio Γ/H changes with time. Freeze-in happens when the system starts in the out-of-equilibrium regime, $\Gamma(T)/H(T) \ll 1$, and moves towards the $\Gamma(T)/H(T) \gg 1$ regime, getting thermalised in the process. Freeze out is the opposite process occurring when an initially thermalized system in the $\Gamma(T)/H(T) \gg 1$ regime moves towards the $\Gamma(T)/H(T) \ll 1$ regime, getting driven out of equilibrium in the process. We will see these processes more explicitly when we study the example below. Before going any further, we would like to point out two caveats of (49) and (50). First, $\Gamma(T)$ here refers to only reactions/collisions that change the number of particles under consideration. Second, (50) is not applicable for unstable, decaying particles.

To get a grasp of the concept of freeze in and freeze out, let us take a look at a gas of electron and positron, interacting mainly via the process $e^+e^- \leftrightarrow \gamma\gamma$[1], in the radiation domination epoch. The

[1] There are of course other processes such as $e^-\gamma \leftrightarrow e^-\gamma$, $e^-e^- \leftrightarrow e^+e^+$, $e^-e^+ \leftrightarrow e^-e^+$, etc, but those processes merely interchange the momenta of the particles involved without changing their particle numbers and therefore irrelevant as far as freeze-in or freeze-out processes are concerned.

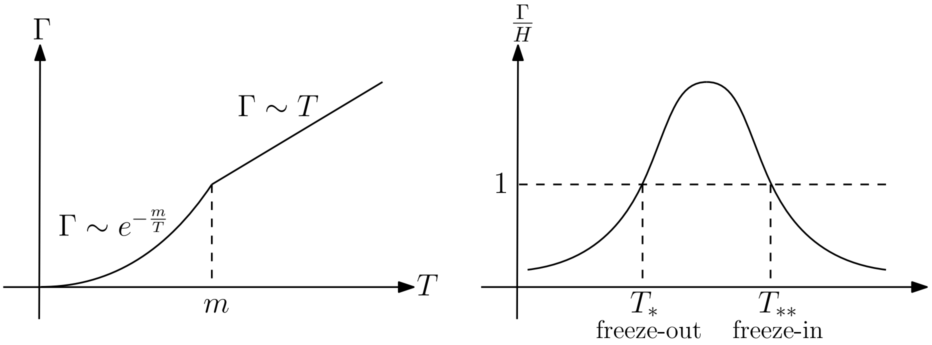


Fig. 2: Temperature dependence of the rate of interaction Γ (left) and its relative size compared to the Hubble parameter Γ/H (right) for the electron-positron-photon system in the radiation domination epoch.

cross-section of the process can be calculated in the centre-of-mass frame to be

$$\sigma \approx \begin{cases} \dfrac{1}{2v}\pi r_e^2, & v \ll 1 \\ \dfrac{m_e^2}{E^2}\pi r_e^2 \left(\log \dfrac{4E^2}{m_e^2} - 1\right), & v \approx 1\,, \end{cases} \tag{51}$$

where $r_e = \alpha/m_e$ with $\alpha = 1/137$ being the fine-structure constant and E is the total energy of the colliding particles in the centre-of-mass frame. The rate of reaction Γ for the cross-section (51) can be computed as

$$\Gamma = \langle \sigma N v \rangle \approx \begin{cases} \pi r_e^2 \left(\dfrac{m_e T}{2\pi}\right)^{3/2} e^{-m_e/T}, & T \lesssim m_e \\ \pi \alpha^2 \dfrac{1}{6}\dfrac{\zeta(3)}{\pi^2} T, & T \gtrsim m_e\,, \end{cases} \tag{52}$$

where we have used (46) in the $T \lesssim m_e$ case, and (41) and $E \sim 3T$ in the $T \gtrsim m_e$ case. The qualitative behaviour of $\Gamma(T)$ is depicted in Figure 2; it is large at high T and small at low T, which follows our intuition: higher temperature implies higher thermal equilibrium density, which, in turn, implies that the particles interact more frequently, hence the higher rate Γ. In the radiation epoch, the Hubble parameter $H(T)$ is given by (44), so

$$\frac{\Gamma(T)}{H(T)} = \begin{cases} \pi r_e^2 \left(\dfrac{m_e T}{2\pi}\right)^{3/2} e^{-m_e/T} \dfrac{M_0}{T^2}, & T \lesssim m_e \\ \pi \alpha^2 \dfrac{1}{6}\dfrac{\zeta(3)}{\pi^2} T \dfrac{M_0}{T^2}, & T \gtrsim m_e\,. \end{cases} \tag{53}$$

As can be seen in Figure 2, $\Gamma(T)/H(T)$ hits one twice, at $T = T_{**}$ (freeze in) and $T = T_*$ (freeze out). These crossing points can be found by equating the above expression to 1. The freeze-in temperature T_{**}, which lies in the $T \gtrsim m_e$ regime, is easier to calculate

$$T_{**} = \alpha^2 \frac{\zeta(3)}{6\pi} M_0 \sim 10^{14}\,\text{GeV}\,. \tag{54}$$

Calculating the freeze-out temperature T_*, which lies in the $T \lesssim m_e$ regime, is somewhat more involved.

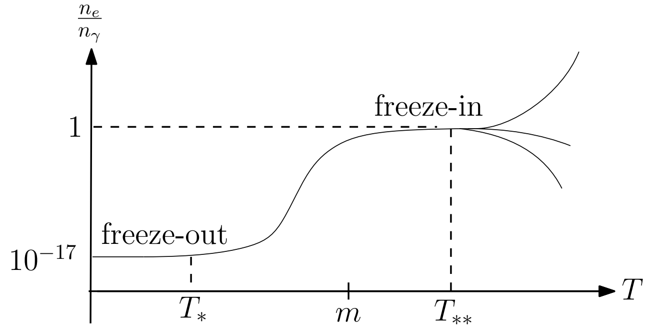


Fig. 3: The behaviour of N_e/N_γ during the freeze-in and freeze-out process. The three different curves on the right part of the graph illustrate the fact that the final, freeze-out value of N_e/N_γ does not depend on the initial value of N_e/N_γ before freeze-in.

It gives

$$\frac{m_e}{T_*} \approx 43\,. \tag{55}$$

After the freeze out, the comoving number density freezes at around its value at $T = T_*$.

In this example, the density of (massless) photons scales as $N_\gamma \propto T^3$ all the time. This means that we can use the ratio N_e/N_γ as a representative of the number density of the electrons. The behaviour of N_e/N_γ during the freeze in and freeze out is shown in Figure. 3. After freeze out ($T < T_*$) N_e/N_γ freezes at its value at $T = T_*$

$$\left.\frac{N_e}{N_\gamma}\right|_{T \lesssim T_*} \sim \left.\frac{N_e}{N_\gamma}\right|_{T=T_*} \approx \left(\frac{43}{2\pi}\right)^{3/2} e^{-43}\,\frac{\pi^2}{\zeta(3)} \sim 10^{-17}\,. \tag{56}$$

Note that the process of freeze-in, which is nothing but thermalisation, erases any information about the initial conditions of electrons in the Universe. Consequently, the freeze-out density of N_e/N_γ does not depend on the initial state of the electrons at $T > T_*$.

There are plenty of phenomena that can be associated with the freeze-in or freeze-out of different interactions. Below we are going to discuss the decoupling of photons, and neutrinos in the early Universe, nucleosynthesis and baryogenesis as examples.

8 Decoupling of photons

If the temperature of the Universe is larger than the binding energy of electrons in atoms, the cosmic plasma is ionized and the mean free path of photons is rather small so that photons are in thermal equilibrium. When the temperature drops, plasma neutralises and the photons no longer interact with matter but propagate freely. The cosmic microwave radiation, which is observed today, is a snapshot of the Universe at the moment of decoupling. Thus, by the study of CMB today we may find the temperature and matter-density fluctuations, existing at redshifts associated with the photon decoupling.

To estimate the temperature of decoupling one notes that the main reactions to be taken into account are the scattering of photons on electrons, $e\gamma \leftrightarrow e\gamma$ (the cross-section of the γp reaction is much smaller) and the reaction of hydrogen (H) dissociation, $ep \leftrightarrow H\gamma$, that controls the concentration of free electrons. When the second reaction is in thermal equilibrium, concentrations of electrons (n_e), protons

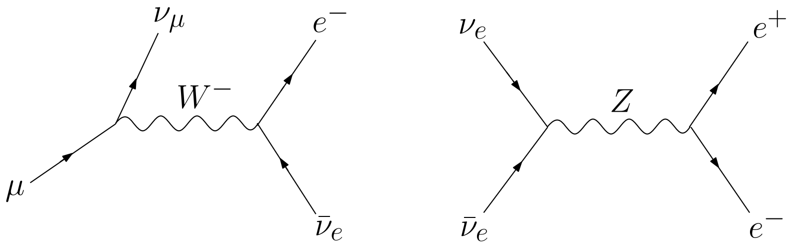


Fig. 4: Charged-current (left) and neutral-current (right) interactions involving neutrinos.

(n_p) and hydrogen atoms (n_H) are related by the Saha formula

$$\frac{n_e n_p}{n_H} = \left(\frac{m_e T}{2\pi}\right)^{\frac{3}{2}} \exp\left(-\frac{I}{T}\right) , \tag{57}$$

where $I = 13.6$ eV $= 1.58 \cdot 10^5$ K is the ionization energy. The decoupling moment is determined by the solution of the equation $\sigma_{\gamma e} n_e \simeq H$, where the Compton scattering cross-section is $\sigma_{\gamma e} = \frac{8\pi\alpha^2}{3m_e^2}$. The system of equations is closed by adding the condition of plasma neutrality ($n_e = n_p$) and introducing as an input parameter the ratio of baryon number (n_B) to the number of photons (n_γ),

$$\eta = \frac{n_B}{n_\gamma} = \frac{n_p + n_H}{n_\gamma} = 6.15 \cdot 10^{-10} , \tag{58}$$

found from different observations.

In addition, the relation between time and temperature can be taken[2] from eq. (44). Numerically, $T^* = 0.25$ eV $= 3000$ K, $z = 1100$, which corresponds to the age of the Universe $t_{\text{dec}} \simeq 3.8 \cdot 10^5$ years.

Above T^* the photons are in thermal equilibrium and thus have the thermal Planck spectrum. Below T^* phonons are decoupled and their moment are red shifted. Since the photons are massless, this redshift is equivalent to the change of temperature, explaining why the CMB observed today has the black-body spectrum.

9 Freeze-out and present concentration of relic neutrinos

Apart from the CMB photons, the present Universe is also filled with relic neutrinos that were once in thermal equilibrium but decoupled at some point. Let us estimate the present concentration of these relic neutrinos. For the current purpose, it is safe to assume that the neutrinos are massless.

Before their decoupling, the neutrinos were held in thermal equilibrium mainly by the charged-current and neutral-current interactions, both of which are depicted in Figure 4. The cross-section for these interactions is roughly

$$\langle \sigma_W v \rangle \sim G_F^2 E^2 , \tag{59}$$

where $G_F = g^2/M_W^2 \sim 10^{-5}\,\text{GeV}^{-2}$ is the Fermi constant. For $E \sim 1\,\text{MeV}$, we have $\sigma_W \sim 10^{-16}$ GeV^{-2}, which is tiny in comparison to typical cross-section for electromagnetic interactions at

[2] Strictly speaking, the decoupling of photons occurs in the epoch of matter dominance, shortly after the moment at which $\Omega_M = \Omega_r$. Accounting for this fact has only a slight influence on the estimate.

the same energy $\sigma_{\text{EM}} \sim \alpha^2/E^2 \sim 10^2$ GeV^{-2}. Neutrino decoupling happens when the rate of reaction

$$\Gamma_W = \langle \sigma_W n v \rangle \sim G_F^2 T^5 \tag{60}$$

is equal to the Hubble parameter (44), namely at

$$T_* \sim \left(G_F^2 M_0\right)^{-1/3} \sim 2\,\text{MeV}\,. \tag{61}$$

Interestingly, T_* is larger than the electron mass but much smaller than the masses of all particles other than photons and neutrinos. It means that at $T \approx T_*$ the Universe is populated by a thermal mixture of e^+, e^-, γ, ν_e, ν_μ, and ν_τ and their antiparticles.

Not long after the neutrino decoupling, the temperature of the Universe goes below electron mass due to the expansion, and e^+'s and e^-'s annihilating into photons. By the time $T \ll m_e$, the remaining particles are γ, ν_e, ν_μ, and ν_τ. The e^+-e^- annihilation process increases the temperature of photons but not that of neutrinos, thus creating a difference between the two temperatures. Let us work out explicitly how much they differ. A time long before the annihilations, when the scale factor is a_{in}, both the photons and neutrinos have a temperature T_{in}. And, long after the annihilations, when the scale factor is a_{out}, the temperature of photons and neutrinos are T_γ and T_ν respectively. Since the Universe expands adiabatically, $H \ll \Gamma_W$, we can make use of the conservation of entropy to determine the change in the photon temperature. During the e^+-e^- annihilation, the entropy of e^+'s and e^-'s, which possess $7/8 \times (2+2)$ degrees of freedom, is transferred to photons, which possess 2 degrees of freedom. Therefore, the entropy conservation for e^-, e^+, and γ reads

$$\begin{aligned} \left[s_{e^+} + s_{e^-} + s_\gamma\right]\big|_{a=a_{\text{in}}} a_{\text{in}}^3 &= s_\gamma\big|_{a=a_{\text{out}}} a_{\text{out}}^3\,, \\ \left[2 + \frac{7}{8}\,(2+2)\right] T_{\text{in}}^3 a_{\text{in}}^3 &= 2T_\gamma^3 a_{\text{out}}^3\,. \end{aligned} \tag{62}$$

Neutrinos, on the other hand, are unaffected by the annihilations, so their temperature simply rescales as

$$T_{\nu,\text{in}} a_{\text{in}} = T_{\nu,\text{out}} a_{\text{out}}\,. \tag{63}$$

Combining (62) and (63), we find the present temperature of relic neutrinos to be

$$T_{\nu,0} = \left(\frac{4}{11}\right)^{1/3} T_{\gamma,0} = 2\,\text{K}\,, \tag{64}$$

where $T_{\gamma,0} = 2.73$ K.

Independently of whether the neutrinos are relativistic or non-relativistic today, we can estimate their present number density. The electron mass, and therefore the temperature at which the electrons and positrons annihilated, is much larger than the upper bounds on the neutrino masses. So, the neutrinos can be assumed to be relativistic close to the time of electron-positron annihilations. We can thus relate the number density of neutrinos with that of photons at a time not long after the annihilations as follows

$$n_\nu = \frac{3}{4}\frac{g_\nu T_\nu^3}{g_\gamma T_\gamma^3} n_\gamma = \frac{3/4}{2}\left(\frac{T_\nu}{T_\gamma}\right)^3 n_\gamma = \frac{3}{22} n_\gamma \,. \tag{65}$$

Since both the number density of neutrino and photon scale in the same way as $n \propto a^{-3}$, the above relation is preserved until today and, for one neutrino degree of freedom,

$$n_{\nu,0} = \frac{3}{22} n_{\gamma,0} \simeq 56\, \frac{1}{\mathrm{cm}^3} \,. \tag{66}$$

This result can be used for getting an upper bound on neutrino masses, by requiring that the total energy density of the neutrinos is smaller than the observed energy density of dark matter. The energy density of dark matter is

$$\rho_{\mathrm{DM}} = \rho_c \Omega_{\mathrm{DM}} \,, \tag{67}$$

where Ω_{DM} is the dark matter abundance and ρ_c is the critical density define in (31). The total energy density of neutrinos today is given by

$$\rho_{\nu,0} = \sum m_\nu n_{\nu,0} = \sum m_\nu \times \left(\frac{3}{22} n_{\gamma,0}\right) \approx \sum m_\nu \times \left(56\ \mathrm{cm}^{-3}\right) \,. \tag{68}$$

In the last step, (64) and the observed number density of CMB photons $n_\gamma = 411\ \mathrm{cm}^{-3}$ was used. Requiring that $2\rho_{\nu,0} < \rho_{\mathrm{DM}}$ (the factor of 2 accounts for antineutrinos), we obtain a very robust bound on the sum of the neutrino masses

$$\sum m_\nu < 100 h^2 \Omega_{\mathrm{DM}}\ \mathrm{eV} \approx 10\ \mathrm{eV} \,, \tag{69}$$

where we have used $h = 0.67$ and $\Omega_{\mathrm{DM}} \approx 0.27$. This bound is more stringent than the direct constraints from particle physics on muon and tau neutrinos. As for experimental detection of the cosmic relic neutrinos, this is very challenging, because of the small interaction cross-section.

10 Big Bang Nucleosynthesis

Roughly speaking, the baryonic matter sector of the Universe consists of 74% of hydrogen and 24% of helium in terms of mass, with a small contribution of other elements. Thermodynamic cooking in stars is not sufficient to explain the observed amount of helium and other light elements in the Universe, suggesting that a cosmological explanation is needed. Big Bang nucleosynthesis (BBN) is our best cosmological explanation for these abundances.

At temperatures above a few hundred MeV, the best description of the Universe is provided by the quark-gluon plasma. Below this temperature, but above a few MeV (the binding energy of protons and neutrons in nuclei) the primordial plasma consists of nucleons rather than nuclei. At smaller temperatures, it is energetically more favourable to hide protons and neutrons in nuclei. The question arises whether all the chemical content of the Universe can be explained by the nuclear reactions occurring at $T \sim 1$ MeV. If not, which elements can be created?

Deviations from thermal equilibrium coming from the expansion of the Universe play an important role in nucleosynthesis. Indeed, in thermal equilibrium, all baryon number would reside in nuclei with the maximal binding energy per nucleon, which is ^{56}Fe. Thus, the dynamics of decoupling of different nuclear reactions must be taken into account. Nuclear abundances are obtained from the solution of a system of kinetic equations incorporating different processes in the expanding Universe. There are various computer codes written for this purpose, which use experimental data for cross-sections of nuclear reactions, supplemented by necessary theoretical information. We shall not discuss this in any detail.

Instead, we will estimate He^4 abundance, which can be done without complicated computations. The first step is to determine the freezing concentration of neutrons. The equilibrium ratio of neutron to proton concentration is simply

$$\frac{n_n}{n_p} = \exp\left(-\frac{m_n - m_p}{T}\right) . \tag{70}$$

It is smaller than unity because neutrons are heavier than protons. The fastest reaction that keeps neutron concentration in equilibrium is $p + e \leftrightarrow \nu + n$. A computation similar to that discussed in Section 9 shows that it goes out of equilibrium at $T \simeq 0.8$ MeV. Therefore, $\frac{n_n}{n_p} \simeq \frac{1}{5}$ for temperature $T \simeq T^*$. This ratio becomes somewhat smaller by the time the BBN starts because of neutron decays, see below.

We can define the starting point of BBN as the point when the first process in the chain reaction $\mathrm{p} + \mathrm{n} \leftrightarrow D + \gamma$ "hides" almost all protons inside the deuterium, i.e. when $n_{\mathrm{p}} \sim n_{\mathrm{D}}$. The relevant Saha's equation in the present case is

$$\frac{n_{\mathrm{p}} n_{\mathrm{n}}}{n_{\mathrm{D}}} = \left(\frac{m_{\mathrm{p}} T}{2\pi}\right)^{3/2} e^{-\Delta_{\mathrm{D}}/T} \tag{71}$$

with $\Delta_{\mathrm{D}} = m_{\mathrm{p}} + m_{\mathrm{n}} - m_{\mathrm{D}} = 2.23$ MeV, where index "D" is short for deuterium. When $n_{\mathrm{p}} \sim n_{\mathrm{D}}$, the Saha's equation reduces to

$$n_{\mathrm{n}} \sim \left(\frac{m_{\mathrm{p}} T}{2\pi}\right)^{3/2} e^{-\Delta_{\mathrm{D}}/T} . \tag{72}$$

Comparing it with (58) and using $n_\gamma \sim T^3$, we find that it occurs when the temperature is

$$T_{\mathrm{BBN}} = 70 \text{ keV} \tag{73}$$

corresponding to the time $t_{\mathrm{BBN}} \approx 4.5$min, which is comparable to the lifetime of neutron. Thus, some portion of neutrons must have decayed by then. By the time the deuterium production becomes effective at T_{BBN}, the ratio of neutron and proton density has reduced from 1/5 to

$$\frac{n_{\mathrm{n}}}{n_{\mathrm{p}}} = \exp\left(-\frac{\Delta_{\mathrm{D}}}{T_*}\right) \exp\left(-\frac{t_{\mathrm{BBN}}}{\tau_{\mathrm{n}}}\right) \approx \frac{1}{7} . \tag{74}$$

Now, if one looks at the binding energies of light elements (say, with an atomic number smaller than 8, the cross-sections for the creation of even heavier elements are exponentially suppressed because of the Coulomb barrier) one finds that it is the highest in He^4. Thus, the abundance Y of He^4 is given simply by the number of available free neutrons in the plasma,

$$Y = \frac{\text{number of baryons in He}^4}{\text{total number of baryons}} = \frac{2n_n}{n_n + n_p} \simeq 0.25 . \tag{75}$$

Abundances of other light elements (He^3, D and Li) can be found from kinetic equations and theoretical predictions can be compared with cosmological observations, see Figure on page 64 of the slides of the Lecture 1. It is plotted as a function of parameter $\eta = \frac{n_B}{n_\gamma} = \eta_{10} \cdot 10^{-10}$, showing the ratio of baryon to photon density for the case of three neutrino species[3]. Amazingly, all light-element abundances are following observations if η is in the interval $\eta = 6.14(19) \times 10^{-10}$, which may be considered as a most important confirmation of the Big Bang theory up to temperatures of the order of 1 MeV. Other elements, present in the Universe, with atomic numbers greater than 12 are believed to be created in massive stars, while lighter elements, such as B, ^{6}Be, and Li are created by a cosmic ray spallation process.

11 Baryogenesis

11.1 Baryon asymmetry in the early Universe

As we discussed in the previous sections, the parameter $\eta = \frac{n_B}{n_\gamma}$, plays an important role in cosmology. It determines the moment of matter-radiation equality and influences primordial abundances of light elements and structure formation. As we shall see, it is related to the fact that there is no antimatter in the Universe (at least, not in amounts comparable to matter).

Let us understand the parameter η better. Once the temperature goes below the electroweak scale, ~ 130 GeV, the baryon number is known to be conserved to a very good approximation (see below). In an expanding Universe, this translates to

$$(n_B - n_{\bar{B}})a^3 = \text{const}\,, \tag{76}$$

where $n_{\bar{B}}$ is the concentration of antibaryons. Furthermore, as long as the Universe expands adiabatically, the entropy is conserved

$$sa^3 = \text{const}\,. \tag{77}$$

If we take the ratio between (76) and (77), the scale factors would cancel and we get

$$\frac{n_B - n_{\bar{B}}}{s} = \text{const}\,, \tag{78}$$

Taking $n_{\bar{B}} = 0$ and $s \simeq 7n_\gamma$ as the values of these quantities today, one gets

$$\frac{n_B - n_{\bar{B}}}{s} \simeq \frac{1}{7}\eta \simeq 9 \times 10^{-11}\,. \tag{79}$$

Consider now high temperatures, say $T \simeq 130$ GeV. The number of quark degrees of freedom at this time is $\simeq 72$, giving $s = 21.4(n_B + n_{\bar{B}})$. This allows us to estimate the baryon asymmetry defined as

$$\Delta \equiv \frac{n_B - n_{\bar{B}}}{n_B + n_{\bar{B}}} \simeq 3\eta \simeq 2 \times 10^{-10} \tag{80}$$

at this temperature. In other words, the baryon-to-photon ratio today gives us an estimate of the (tiny) baryon asymmetry in the early Universe.

[3] Changing the number of neutrino species changes the rate of the Universal expansion and thus predictions of Big Bang nucleosynthesis. One cannot admit more than four types of massless neutrinos in order not to spoil successful predictions of BBN.

When the Universe cools down from this state, the symmetric part of the baryon-antibaryon background annihilates into photons and neutrinos, but the nucleons that do not find a pair survive, see Fig. 5. These give rise to baryonic matter in the Universe, and eventually galaxies, stars and planets.

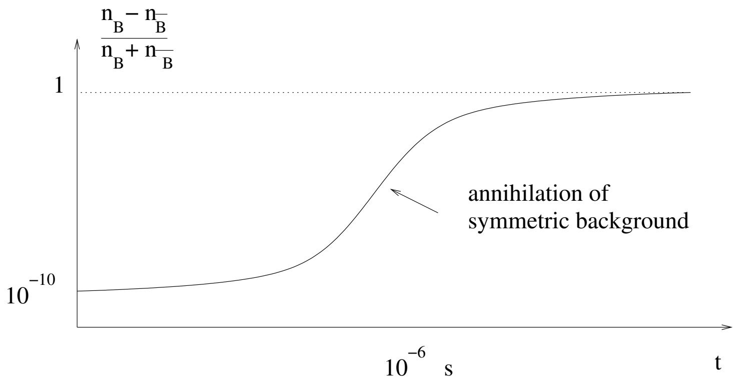


Fig. 5: Dependence of baryon asymmetry on time.

11.2 Sakharov Conditions

Rather than assume that the Universe contained more baryons than antibaryons from the very beginning, it is more compelling to think that the Universe started in a baryon-symmetric state and somehow produced baryon asymmetry as it evolved. The initial baryon symmetric state may be a consequence of cosmological inflation, as will be discussed below. The process of creating baryon asymmetry is dubbed baryogenesis.

From the viewpoint of particle physics, to produce a net baryon asymmetry in the Universe the three so-called Sakharov conditions must be met:

1. Existence of baryon number violating processes.
2. Violation of discreet symmetries C and CP.
3. Departure from thermal equilibrium.

In this case, the Universe could start its expansion from a truly symmetric state, containing an equal number of particles and antiparticles. Then, in the course of the expansion, the particle physics reactions with B, C and CP non-conservation would produce an excess of particles over antiparticles. The third condition is required because in thermal equilibrium the baryon number of the system must be zero: the total rate of the processes which increase the baryon number is exactly compensated by the rate of the processes that decrease it, as a consequence of the CPT theorem.

Depending on the way the three Sakharov conditions are implemented, one can consider different types of baryogenesis mechanisms. We will briefly review two of the most popular ideas, called leptogenesis and domain wall electroweak baryogenesis (for detailed reviews see, e.g. [7–10]), leaving others (such as Grand Unified baryogenesis, Affleck-Dine baryogenesis, baryogenesis from black holes, spontaneous baryogenesis, etc) for individual study (for reviews see, e.g. [11–13]).

11.2.1 Baryon number violation in the Standard Model

Qualitatively speaking, the Standard Model has all the necessary ingredients to create baryon asymmetry: the Cabibbo-Kobayashi-Maskawa (CKM) matrix that describes the mixing among the different quark flavours contains a CP-violating phase, a departure from thermal equilibrium comes from the Universe expansion, and baryon-number violation occurs through the so-called "sphaleron" processes.

On a perturbative level, the Standard Model has four conserved fermionic numbers: B - baryon number, and $L_e,\ L_\mu,\ L_\tau$ - leptonic numbers. The non-perturbative effects break these conservation laws and leave intact only three of them, $L_e - B/3,\ L_\mu - B/3$, and $L_\tau - B/3$. In other words, there is anomalous fermion-number non-conservation in the SM. It can not be described by Feynman diagrams and uses for its description several advanced non-perturbative methods of Quantum Field Theory, such as triangular anomalies, instantons, and non-trivial topological structure of the gauge vacua.

The sphaleron process is a non-perturbative high-temperature reaction violating (B+L) that does roughly the following conversion:

$$9q + 3l \to \mathcal{O}\left(\frac{1}{\alpha_W}\right) bosons\,.$$

It requires the presence of a large number of bosons, coming out from a "sphaleron" decay, with sphaleron being a specific static but unstable solution of the classical equations of motion of the bosonic sector of the SM.

The rate of anomalous fermion-number non-conservation at zero and non-zero temperatures is of the order of:

$$\Gamma \sim \begin{cases} \exp(-\frac{4\pi}{\alpha_W}) \sim 10^{-160}, & T = 0 \\ \\ (\alpha_W T)^4 (\frac{M_{sph}}{T})^7 \exp\left(-\frac{M_{sph}}{T}\right), & T < T_c \\ \\ (\alpha_W)^5 T^4, & T > T_c \end{cases} \tag{81}$$

where $M_{sph} \sim M_W/\alpha_W$ is the sphaleron mass and $T_c \simeq 160$ GeV is the temperature of the electroweak (EW) crossover (see below).

The anomalous fermion number violation of the SM looks to be specially designed for baryogenesis: it freezes in at $T \simeq \alpha_W^5 M_0 \sim 10^{12}$ GeV, freezes-out out at $T \simeq 130$ GeV, and is consistent with all constraints on the matter stability at zero temperature, such as the proton decay.

11.3 Baryon asymmetry and the Standard Model

Still, the Standard Model cannot explain the observed baryon asymmetry. The reason is twofold. First, the SM CP violation has a peculiar structure: it vanishes if any pair of up or down-type quarks degenerate in mass. This allows us to estimate the CP-violating effects without a complicated computation. The relevant energy scale for baryogenesis is the temperature at which sphaleron transitions are active, $T > T_{\rm sph}$. This temperature is larger than the quark masses, which can be considered as a small perturbation.

Now, a polynomial constructed from the quark masses, that obeys the above property is

$$P = (m_t^2 - m_c^2)(m_t^2 - m_u^2)(m_c^2 - m_u^2)(m_b^2 - m_s^2)(m_b^2 - m_d^2)(m_s^2 - m_d^2)\,. \tag{82}$$

Note that the squares of the quark mass should be used, as no physical quantity can depend on the first power of the quark mass. To get a measure of CP violation, this combination should be multiplied by $\sin(\theta_{12})\sin(\theta_{23})\sin(\theta_{13})\sin\delta_{CP}$, where δ_{CP} is the KM phase and the Particle Data Group parametrization of the CKM matrix is used. Indeed, if one or more of the mixing angles in the CKM matrix vanishes, then with a physically unobservable change of phases of the quark fields, the CKM matrix can be made purely real and there is no CP violation. Finally, to get a dimensionless quantity, the combination (82) should be divided by the 12th power of the relevant temperature. Numerically, the effects of the Kobayashi-Maskawa CP violation are of the order

$$\delta_{KM}^{CP} \sim \frac{P}{T^{12}} \sim 10^{-20}\,, \tag{83}$$

i.e., some ten orders of magnitude smaller than the baryon asymmetry of the Universe.

Secondly, the rates of strong, weak and electromagnetic interactions at the sphaleron freeze-out temperature 130 GeV are much faster than the Hubble rate. They keep the distributions of quarks and leptons in equilibrium, and, because of this, the production of any sizable baryon asymmetry is greatly suppressed, as follows from the third Sakharov condition. So, the presence of the baryon asymmetry of the Universe is a strong argument in favour of new physics, which should bring new sources of CP-violation and new mechanisms for the departure from thermal equilibrium.

11.3.1 Leptogenesis

The Standard Model (SM) of elementary particles, defined as a renormalizable field theory, based on the SU(3)×SU(2)×U(1) gauge group, and containing three fermionic families with left-handed particles being the SU(2) doublets, the right-handed ones being the SU(2) singlets and one Higgs doublet, has been used to successfully predict several particles and their properties. It has a strange asymmetry between the quark and lepton sectors. Namely, every left-handed quark or charged lepton has its right-handed counterpart. The neutral leptons – neutrinos are in the SM in different ways – they do not have right-handed partners. This is exactly the reason why the neutrinos are exactly massless in the SM and why there are three distinct conserved leptonic numbers. Nowadays, it is well established experimentally that neutrinos change their flavours and thus have nonzero masses. Therefore, it is natural to upgrade the SM by adding to it $\mathcal{N} = 3$ right-handed neutral fermion states. They can be called singlet or neutral leptons, sterile neutrinos, or alike. The PDG name for them is “heavy neutral lepton”, HNL for short. The number “3” looks natural as this is the number of fermionic generations in the SM. It also happens to be a minimal number, which enables us to explain simultaneously neutrino masses and oscillations, and the presence of baryon asymmetry and Dark Matter (DM) in the Universe.

The most general renormalisable Lagrangian including the SM and several right-handed singlet fermions N_I reads

$$L = L_{SM} + \bar{N}_I i\partial_\mu \gamma^\mu N_I - F_{\alpha I}\, \bar{L}_\alpha N_I \tilde{\phi} - \frac{M_I}{2}\, \bar{N}_I^c N_I + h.c., \tag{84}$$

where ϕ is the Higgs field and $\tilde{\phi}_i = \epsilon_{ij}\phi_j^*$. This Lagrangian can be used for the explanation of the small values of neutrino masses via the see-saw mechanism, see the lectures on neutrino physics at this School by Gabriela Barenboim. If the masses of HNLs are much larger than the electroweak scale, say of the order of 10^{10}, this Lagrangian is called the “type I see-saw model”. If the masses of HNLs are below the Fermi scale, this is dubbed as “νMSM” for the Neutrino Minimal Standard Model, stressing that new fermions are similar to other fermions of the SM (for a review see, e.g. [14]).

A new qualitative feature of this theory is lepton number non-conservation. Indeed, since N_I transform as singlets to the SM gauge group, the Majorana mass terms for them are allowed by the gauge symmetries. The Yukawa interactions of HNLs with the leptons contain CP-violating phases, whereas the mere presence of new particles creates new sources for departure from thermal equilibrium during the freeze in or freeze out of HNLs. HNLs can decay or scatter, producing lepton asymmetry. If this happens above the sphaleron freeze-out, this asymmetry is converted into baryon asymmetry by sphalerons (Section 11.2.1). The resulting baryon asymmetry is just a numerical factor of order one smaller than the generated lepton asymmetry. If the HNL masses are smaller than the mass of the Z-boson, the see-saw mechanism of neutrino mass generation and baryogenesis can be verified experimentally at new experiments at CERN such as SHiP, and later at FCC-ee, see Figure on page 16 of the Lecture 2 slides.

11.3.2 Phase transitions and baryogenesis

Yet another idea of how to produce baryon asymmetry in the Universe is associated with first-order phase transitions. The same substance, depending on temperature T and pressure P, can be in a different phase state. For instance, the water can be in solid, liquid and gaseous states, depending on P and T. Similarly, a hot dense plasma of elementary particles can be in different phases. For instance, there could be a first-order phase transition between the symmetric and Higgs phases of the Standard Model when the Universe was cooling down.

The electroweak phase transition (EWPT), if it is strongly first order, would be quite a violent event. The first-order phase transitions go through the bubble nucleation, qualitatively in the same way the water boils or the vapour condenses. EWPT would start at $T_c \sim 100$ GeV. At this time the size of the event horizon is $\simeq 1$ cm. The critical bubbles are of microscopical size, roughly $R \sim (\alpha_W T)^{-1}$. Bubbles nucleate in different places, expand, and eventually collide, filling the Universe with a new phase. The typical size of bubbles at the time they collide is macroscopic, $\sim 10^{-6}$ cm. Inside the bubble the system is in the Higgs phase and the vacuum expectation value of the Higgs field is non-zero. Outside the bubble, the system is in the symmetric phase. Inside the bubble, the rate of B-violation must be small to evade the destruction of the baryon number due to sphalerons. This requirement puts a constraint on the strength of the phase transition. In contrast, outside the bubble the rate is large. The masses of particles are different inside and outside the bubbles (particles inside the bubble are generically heavier since they get their masses from interactions with a Higgs field), leading to their interaction with the bubble walls.

The non-equilibrium motion of the bubble walls, combined with the effects of CP-violation of

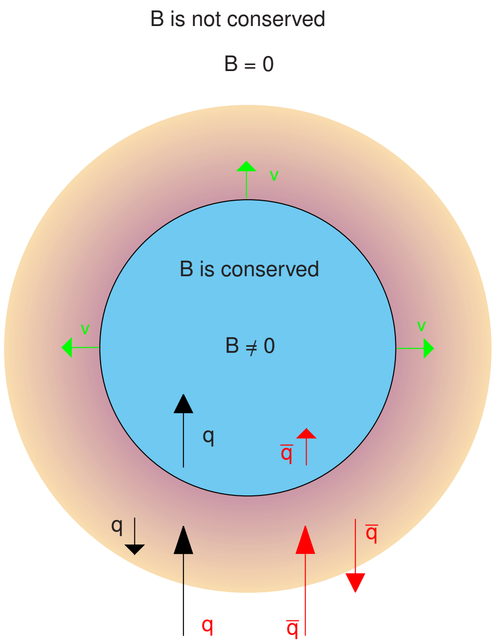


Fig. 6: The expanding bubble. The dark region inside the small circle corresponds to the broken (Higgs) phase where quarks and leptons are massive and baryon number violating processes are strongly suppressed. The region between the two circles corresponds to the plasma in the symmetric phase which is disturbed by the motion of the bubble wall. It is this region which is responsible for the generation of baryon excess. The rate of baryon number non-conservation is high here, while CP-non-invariant interaction of fermions with the domain wall spatially separates fermions from antifermions in a way that fermions are going inside the bubble. The region outside the large circle corresponds to the undisturbed symmetric phase.

fermion scattering on the walls, leads to the separation of baryon number: fermions go inside the bubbles and antifermions outside. The outside excess of antifermions is destroyed in equilibrium sphaleron reactions, whereas the excess of fermions in the Higgs phase remains intact. As the region of the broken phase increases, at the time when different bubbles collide, the whole Universe is baryon-asymmetric.

The scenario described above does not work for the SM. The phase diagram for the electroweak theory is shown in Fig. 7. The vertical axis is the temperature, while the horizontal axis is the Higgs mass in the SM. The phase diagram in many extensions of the standard model looks qualitatively the same, but the horizontal axis for them is a combination of the different parameters rather than the Higgs mass itself.

This diagram is similar to that of a liquid–vapour system. There is an end-point of a line of the first-order phase transitions. If the Higgs mass is equal to the critical value, the phase transition in the system is of second order. For smaller Higgs masses the EW phase transition is of the first kind, while at larger Higgs masses there is no phase transition at all.

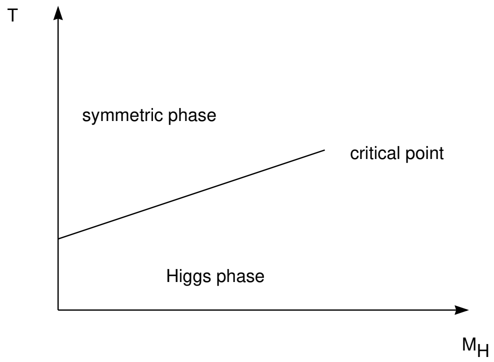


Fig. 7: The phase diagram of the electroweak theory. For small Higgs masses, to the left from the critical point, the electroweak phase transition is of the first kind. At the critical point, it is of the second kind, and the critical properties of the electroweak theory near this point are very similar to those of the liquid-vapour system. At large Higgs masses, to the right from the critical point, there is no phase transition at all.

For the SM, the endpoint of the first order phase transition line can be computed unambiguously and corresponds to $M_H \simeq 72$ GeV. With the experimental value of the Higgs mass 125 GeV, there is no phase transition at all. So, the SM can be excluded as a theory for domain wall electroweak baryogenesis. Extensions of the SM may be viable from the point of view of the strength of the phase transition. For instance, extra scalars in the mass range $\sim$ 100 GeV may lead to the necessary strength of the phase transition.

Assume now that we have a model with a necessary first-order phase transition. What is the *magnitude* of the baryon asymmetry created? It depends on the particle content, the strength of CP-violation, the rate of B-non-conservation in the symmetric phase, the bubble wall velocity, and many other details of interactions. Computations of baryon excess have been carried out in different models, with the result that the asymmetry 10^{-10} may be derived, provided the parameters of a model are chosen in some particular way. In addition to constraints on the particle spectrum, following the requirement of a first-order phase transition, there appear extra constraints to the models, since a sufficient amount of CP violation is required.

12 Dark Matter

12.1 Evidence for Dark Matter

The existence of non-luminous and non-absorbing substance, dubbed dark matter, making up about 27% of the mass of the Universe is now an established fact. Our confidence in its existence has been built on many independent observations pointing to the same conclusion. Let us discuss some of them.

Rotational curves of spiral galaxies. Compelling evidence for dark matter is coming from the orbital velocities of stars located on the disk of a spiral galaxy. The mass distribution of luminous matter in such a galaxy can be inferred by measuring its luminosity as a function of radius. Observations show that the luminosity follows a rough radial dependence of the form

$$I(r) = I_0 \exp\left(-\frac{r}{r_0}\right) . \tag{85}$$

From here we can infer that most of the mass in a spiral galaxy is concentrated near its centre. If we assume that only luminous matter gravitates, the velocities of the stars on the disk at the outer side of the galaxy would fall off according to Kepler's law ($v \propto 1/\sqrt{r}$). Indeed, since most of the mass is clumped near the centre of the galaxy, stars located at different radii sufficiently far from the centre of the galaxy feel essentially the same amount of mass pulling them towards the centre. Mathematically,

$$mv^2 \approx G\frac{Mm}{r} \Longrightarrow v \approx \sqrt{\frac{GM}{r}} . \tag{86}$$

In reality, what we see is a flat profile

$$v \approx \text{constant} \tag{87}$$

suggesting that the M in (86) is not a constant but varies as

$$M(r) \propto r \tag{88}$$

which could be explained by the presence of a dark matter halo with mass distribution at large r

$$\rho_{\text{dark}}(r) \propto \frac{1}{r^2} . \tag{89}$$

Gravitational lensing. Gravitational lensing is a phenomenon where massive objects located between us and distant light sources act as a lens, bending the space around them and consequently bending any light passing nearby. When the effect only causes slight shear deformations in the images of distant luminous objects, it is called weak gravitational lensing. By measuring the amount of such shear deformations on distant galaxies, and combining it with appropriate statistical analysis, we can infer the total mass of the intervening galaxy clusters causing the deformations. This led to the same conclusion that the galaxy clusters are more massive than the total mass of the visible objects belonging to them.

Bullet Cluster. The Bullet Cluster is an aftermath of a collision between two galaxy clusters. The mass distribution of the colliding clusters can be inferred by gravitational lensing. It was found that the non-luminous matter simply passes through one another, showing that the dark matter has essentially no pressure. Furthermore, it was observed that the location of the centre of mass inferred from gravitational lensing is significantly displaced from that of the (luminous) baryonic matter, a fact that can be explained easily by the presence of dark matter.

Big Bang nucleosynthesis. The BBN predicts the abundance of light elements in the Universe as a function of the baryon-to-photon ratio η. Requiring the predictions to match with the observed values allows us to determine the baryon abundance $\Omega_B \simeq 0.05$. Numerically Ω_B is smaller than the observed abundance Ω_M of non-relativistic matter in the Universe, which also suggests the existence of dark matter.

Combined data from SNe, BAO, and CMB. The combined data from supernovae (SNe), baryon acoustic oscillation (BAO), and the cosmic microwave background (CMB) constraint the abundances of non-relativistic matter Ω_M, non-relativistic matter in the form of baryons Ω_B, and the cosmological constant Ω_Λ to have the values in Equation (35). The fact that $\Omega_B < \Omega_M$ by a significant amount suggest the presence of non-baryonic matter, i.e. dark matter.

12.2 Constraints on dark matter particle

The SM does not provide any candidate for the non-baryonic dark matter and, therefore, cosmological observations point in the direction of physics beyond the standard model.

Lower bound on the mass of fermion DM particle. The Pauli exclusion principle gives an upper limit on how densely fermions can be packed in the phase space $(\mathbf{p}, \mathbf{x})$. Consequently, the number of fermions N_F with typical velocity v contained within an object of size r, e.g. a galaxy halo, is bounded from above

$$N_F \lesssim \frac{1}{(2\pi)^3} \int d^3\mathbf{p} d^3\mathbf{x} n \sim \left(\frac{pr}{2\pi}\right)^3 . \tag{90}$$

In the last step, we have assumed that n has the form of a window function with momentum and coordinate extent of p and r respectively. It follows that the total mass of fermions in a galaxy M is also bounded from above

$$M \lesssim m_\nu N_F \sim m_F^4 (vr/2\pi)^3 . \tag{91}$$

At the same time, Kepler's law states that $v^2 \sim GM/r$, which, when combined with the above bound, gives a lower limit on the masses of fermions in terms of measurable quantities v and r

$$m_F \gtrsim \left(\frac{(2\pi)^3}{Gvr^2}\right)^{1/4} \approx 120 \left(\frac{100\ \text{km/s}}{v}\right)^{1/4} \left(\frac{1\ \text{kpc}}{r}\right)^{1/2} . \tag{92}$$

This constraint is known as the Tremaire-Gunn limit. For example, in our galaxy $r \sim 10$ kpc and $v \sim 220$ km/s, giving the fermion mass bound of $m_F \gtrsim 30$ eV. A more stringent bound is given by dwarf spheroidal galaxies whose typical masses are around $M \sim 10^6 M_\odot$. Those galaxies give the fermion mass bounds of $m_F \gtrsim 500$ eV. This bound, in particular, excludes ordinary neutrinos as dark matter candidates.

Lower bound on the mass of boson DM particle. The above argument does not work for bosons, as the number of bosons per one quantum state is not bounded. Still, the de Broglier wavelength $2\pi/(m_B v)$ of a DM particle with escape velocity v and mass m_B must be smaller than the size r of the compact DM object, such as a dwarf galaxy. This leads to a lower bound $m_B > 10^{-22}$ eV. The DM so light is usually called "fuzzy" dark matter.

Cold, warm and hot dark matter. An important constraint on DM particles comes from the study of the structure formation. The analysis of the evolution of the density perturbations reveals that the structure formation starts at the matter radiation equality, at $T_{\text{eq}} \simeq 0.8$ eV, corresponding to the redshift $z \simeq 3500$. The free-streaming length λ_{FS} of the DM particles at this time determines the minimal mass of the structure that can be formed by the gravitational Jeans instabilities,

$$M \lesssim M_{\mathrm{FS}} \simeq \frac{4}{3}\pi\lambda_{\mathrm{FS}}^3\rho_{\mathrm{DM}}\,, \tag{93}$$

where the DM energy density ρ_{DM} is taken at T_{eq}. The DM particle traverses the distance l after its decoupling determined from the equation

$$\frac{dl}{dt} = v(t) + Hl\,, \tag{94}$$

where $v(t)$ is the velocity of DM particle, and H is the Hubble rate. A relatively simple computation leads to

$$M_{\mathrm{FS}} \simeq 10^{10} M_{\odot} \left(\frac{7\ \mathrm{keV}}{M}\right)^3 \frac{g^*(\mathrm{now})}{g^*(\mathrm{decoupling})} \left(\frac{\langle p\rangle}{p_T}\right)^3\,, \tag{95}$$

where $M_{\odot}$ is the solar mass, $g^*(\mathrm{now})$ and $g^*(\mathrm{decoupling})$ are the number of effectively massless degrees of freedom now and at the moment of DM particle decoupling, $\langle p\rangle$ is the average momentum of the DM particle at decoupling, and $p_T \simeq 3.15T$ is the average thermal momentum of a fermion at this time. Equation (95) tells, in particular, that the dark matter particles must not be relativistic at the onset of structure formation, to admit the formation of structures such as dwarf galaxies.

The ordinary neutrino would be a hot DM particle, with the free streaming mass exceeding the mass of a cluster of galaxies $> 10^{14}M_{\odot}$. In this case, the smaller structures could not be formed, excluding once more the ordinary neutrino as a DM candidate. A WIMP (weakly interacting massive particle) with a mass, say, 100 GeV leads to a negligibly small free streaming mass, allowing the formation of structures at small scales. This a cold DM candidate. A 7 keV particle, lying somewhere between these two extremes, would be a "warm" DM matter candidate.

Known and unknown properties of dark matter. To summarise, the dark matter particles must be sufficiently stable with a lifetime that is longer than the age of the Universe, or otherwise, they would have decayed. If the dark matter particles are relatively light, $M \lesssim 1$ TeV, then they must be neutral and very weakly interacting, otherwise, we would have detected them. If the dark matter particles are fermions, their mass must be above 500eV (the Tremaine-Gunn bound we obtained above). If they are bosons, their mass should be above 10^{-22} eV. Also, the free streaming length of DM particles should be small enough in order not to get in conflict with structure formation.

The predictions of dark matter masses from different models vary wildly, they go from as light as 10^{-22} eV (stringy axions) to as heavy as 10^{24} GeV (such as supersymmetric Q-balls). The spin of the dark matter particle is not known. Since the presence of dark matter has been based entirely on its gravitational effects, not much is known about its non-gravitational interactions other than they must be very weak. For the same reason, the production mechanism of dark matter and how they are embedded in the big picture of particle physics are the subjects of speculation.

12.3 Sterile neutrino dark matter

It became fashionable to make vote at different conferences and workshops in favour of this or that DM candidate. At the moment, the first place is usually attributed to WIMP (weakly interacting massive particle), the second to axion, and the third to sterile neutrino. The motivation for WIMPs stems from low-energy supersymmetry and, more generally, from naturalness and the hierarchy problem. The motivation for axion stems from the strong CP problem. The motivation for sterile neutrino comes from neutrino physics. The first two candidates were discussed in the special lecture on Dark Matter by Mads Frandsen (see also the non-shown slides 44-57 of my Lecture 2). In these lecture notes, we will only cover the third one.

Let us consider the theory described in Section 11.3.1, which is just the SM extended by 3 HNLs. As we have already explained, it can describe neutrino masses and oscillations and produce baryon asymmetry of the Universe. In addition, it also gives a suitable DM candidate. Though the νMSM does not have any extra stable particle in comparison with the SM, the lightest singlet fermion, N_1, may play the role of a dark matter particle as its lifetime can greatly exceed the age of the Universe (for a review see, e.g. [15]).

The following considerations determine a possible range of masses and couplings of the DM sterile neutrino:

- The sterile neutrino N_1 can decay via the mixing with the ordinary neutrino into 3 neutrinos, $N_1 \to \nu\nu\bar{\nu},\ \ N_1 \to \bar{\nu}\bar{\nu}\nu$ with the width $\Gamma \propto G_F^2 m_N^2 \theta^2$ where m_N is the mass of the DM sterile neutrino, and $\theta = fv/m_N$ is the mixing angle. Here f is the Yukawa coupling of the sterile neutrino and v is the Higgs vacuum expectation value. The sterile neutrino lifetime must exceed the age of the Universe.
- Cosmological production. DM sterile neutrinos are produced in reactions like $l^+l^- \to N_1\nu$ at temperatures of the order of 100 MeV. If N_1 interact too strongly, these reactions would overproduce them, making the abundance of Dark Matter in the Universe larger than observed. This leads to an upper limit on the strength of interaction of N_1. If N_1 is produced only in the reactions of this type, the requirement to produce enough Dark Matter also results in the lower bound on the mixing angle. This lower bound depends on the conditions in the early Universe during the epoch of N_1 production [14], such as the lepton asymmetry of the Universe. Moreover, the lower bound completely disappears if N_1 can also be produced at very high temperatures by interactions related to gravity or at the end of cosmological inflation.
- X-rays. N_1 decays radiatively with the width $\Gamma_\gamma \propto \alpha G_F^2 m_N^2 \theta^2$, $N_1 \to \gamma\nu$, producing a narrow line that can be detected by X-ray telescopes (such as XMM-Newton or Chandra). These considerations result in an upper limit on the sterile neutrino mixing angle with active neutrinos. While this upper limit depends on the uncertainties in the distribution of dark matter in the Milky Way and the other nearby galaxies and clusters and the modelling of the diffuse X-ray background, it is possible to conservatively marginalise over these uncertainties, obtaining very robust constraints (see [15] and references therein).
- Structure formation. If N_1 is too light, a very large number density of such particles is required to make an observed halo of a small galaxy. As HNLs are fermions, their number density can not exceed that of a completely degenerate Fermi gas, resulting in a very robust lower bound $\simeq 0.5$ keV on the mass of the particle. This bound can be further improved by taking into account that, as discussed above, light-dark matter particles remain relativistic till quite late epochs and therefore

suppress or erase density perturbation on small scales. This would affect the inner structure of the halos of the Milky Way and other galaxies, as well as matter distribution in the intergalactic medium.

The summary of constraints is presented in Figure on slide 43 of Lecture 2. There the solid lines represent the most important constraints that are largely model-independent. The *phase space bound* (solid purple line) is based on Pauli's exclusion principle applied to dark matter in dwarf galaxies. The bounds based on the *non-observation of X-rays* from the decay $N_1 \to \nu\gamma$ are shown by the violet area. The dashed lines represent the estimates of the sensitivity of the future X-ray mission ATHENA. The blue square marks the interpretation of the 3.5 keV excess as decaying sterile neutrino DM. Above the line marked "thermal overproduction", the abundance of sterile neutrinos would exceed the observed DM density. All other constraints depend on the sterile neutrino production mechanism.

The upper limits on the strength of interaction of sterile neutrino allow us to fix the overall scale of active neutrino masses in the νMSM. The DM sterile neutrino effectively decouples from the see-saw formula, telling that the mass of one of the active neutrinos is much smaller than the observed solar and atmospheric mass differences. This fixes the masses of two other active neutrinos to $\simeq 0.009$ eV and $\simeq 0.05$ eV for the normal ordering and to the near degenerate value 0.05 eV for the inverted ordering.

The most promising way to find DM matter neutrino is indirect detection with the use of X-ray telescopes in Space. The new X-ray spectrometer XRISM, which was launched in September 2013, has great potential to detect a signal from Dark Matter decay.

13 Dark Energy

The cosmological observations suggest that the Universe is accelerating now. This is perfectly consistent with a non-zero positive cosmological constant, $\Omega_\Lambda \simeq 0.7$, though maybe a signal of the presence of some exotic substance – Dark Energy – with the equation of state close to that of the cosmological constant, namely $p = \omega\rho$, where experimentally $\omega = -1 \pm 0.1 \pm 0.1$ with the first error being systematic and the second being statistical. The observed value of Ω_Λ corresponds to the vacuum energy density

$$V_0 = \left(\frac{\Lambda}{8\pi G}\right)^{\frac{1}{4}} \sim 2 \times 10^{-3}\,\text{eV} \sim 0.01\,\text{cm}^{-1}\,. \tag{96}$$

The value of the cosmological constant in the SM plus gravity cannot be predicted, in the exact similarity with the other parameters of the SM, such as the mass of electrons or other particles. At the present state of knowledge, all these parameters are simply taken from the experiment. A naive computation of the vacuum energy density in the SM gives a divergent result without physical meaning, representing a sum of zero-point energies of all species of the SM.

A particular value of the cosmological constant poses several questions which remain without the answers. The first comes from a comparison between the magnitude of the scale (96) and other known scales, for example, Λ_{QCD}, M_W, and M_{Pl}. Why is the scale associated with Λ so small compared with the other scales? Another problem arises from the comparison of Ω_M and Ω_Λ. At the present stage of expansion of the Universe, they are of the same order of magnitude. Is this a pure coincidence or there is a deep reason behind that?

14 Inflation

14.1 Problems of standard cosmology

To explain what kind of problems the standard cosmology faced before the invention of inflation, we will introduce the notion of particle horizons. In a static Universe, if two events are separated by distance Δl and time Δt, they are causally independent, provided $\Delta l > \Delta t$. What is the analogue of this statement in an expanding Universe? To answer this, let us write the equation describing the propagation of light, taking into account the fact that the speed of light is one in the natural system of units:

$$\frac{dl}{dt} = 1 + \frac{\dot{a}}{a} l \,. \tag{97}$$

The solution of this equation is

$$l(t) = \int_{t_0}^{t} \frac{a(t)}{a(t')} dt' \,. \tag{98}$$

For both radiation- and matter-dominated Universes, with $a(t) \sim t^{1/2}$ and $a(t) \sim t^{2/3}$ respectively, the integral in eq. (98) converges even if $t_0 = 0$. The distance light travels since t_0 is called the particle horizon, $l_H(t)$. For different epochs, we have:

$$l_H(t) = \begin{cases} 2t, & \text{radiation-dominated epoch,} \\ 3t, & \text{matter-dominated epoch.} \end{cases} \tag{99}$$

If the distance between two points is greater than $l_H(t)$, the points were not in causal contact in the past and thus we should expect that the parameters of the Universe (such as temperature) should be different.

The horizon and homogeneity problem. As we have already discussed, the photons decoupled from the plasma at the redshift $z \simeq 1100$ corresponding to the time t_d. Let us assume that there were only radiation- and matter-dominated epochs in the past. Then, when looking at different points of the sky separated by some angle $\theta > \theta_H$ we would observe CMB emitted from regions that were never in causal contact and so should have different temperatures. To estimate θ_H, one should find the present size L of the region that was the horizon at the decoupling time,

$$L \sim 3t_d \frac{a_{\text{now}}}{a_d} \sim 3t_d \left(\frac{t_{\text{now}}}{t_d}\right)^{\frac{2}{3}} \,, \tag{100}$$

where $3t_d$ is the horizon scale at t_d. The angle θ_H is simply the ratio of this scale to the present size of the horizon,

$$\theta_H \simeq \frac{3t_d}{3t_{\text{now}}} \left(\frac{t_{\text{now}}}{t_d}\right)^{\frac{2}{3}} \simeq \left(\frac{t_d}{t_{\text{now}}}\right)^{\frac{1}{3}} \,, \tag{101}$$

which corresponds to $\theta_H \sim 2°$. This means that the present horizon contains $O\left(\frac{t_{\text{now}}}{t_d}\right) \sim 10^4$ domains that were not in causal contact before recombination, see Fig. 8. However, observations show that the cosmic microwave background is isotropic for these angles, with accuracy better than 10^{-4}. This is the essence of the horizon and homogeneity problem.

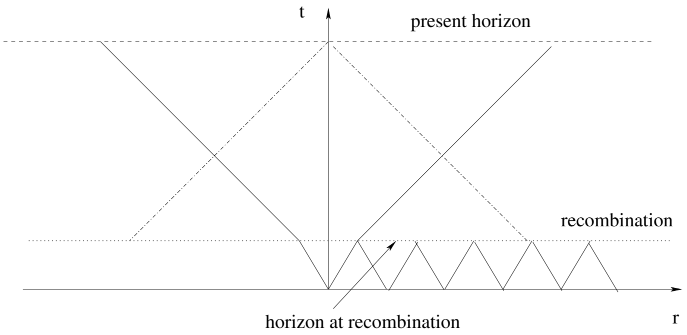


Fig. 8: An observer at point $r = 0$ sees many particle horizons corresponding to photon decoupling.

The flatness problem. Let us consider the relationship (34) in somewhat different form,

$$\Omega - 1 = \frac{k}{a^2 H^2} , \tag{102}$$

where Ω is the ratio of the total energy density to the critical density. For both matter- and radiation-dominated epochs, $H \sim \frac{1}{t}$ and $a \sim t^\alpha$, with $\alpha = \frac{1}{2}$ or $\alpha = \frac{2}{3}$. Thus, Ω increases with t as $\Omega - 1 \sim t^{2(1-\alpha)}$. Therefore, to have $\Omega \simeq 1$ at present, as follows from observation, Ω must have been finely tuned to one with huge accuracy in the past. For example, at the nucleosynthesis time, $|\Omega - 1|$ must be of the order of 10^{-15}. It is unclear why the Universe should have been so flat in the past.

To summarise, the observational fact that the present Universe is flat, homogeneous and isotropic is very bizarre. **If** the Universe was dominated by matter or radiation in the past, then the initial conditions for expansion must be highly fine-tuned. The Universe must have been much flatter than it is now, and the causally disconnected regions must have had the same characteristics such as densities, temperatures, etc. The question arises is there any rationale behind these fine-tuned initial conditions? A possible answer, associated with cosmological inflation, is discussed below.

14.2 Inflation as a solution of cosmological problems

The three problems described above are related to each other. The inflationary paradigm provides a simultaneous solution to all of them.

First, let us note that the assumption that the Universe was dominated by radiation or by matter well in the past does not follow from any observation. We can only be sure that the Universe was dominated by radiation at the epoch of BBN, perhaps at the time of baryogenesis at $T \sim 100$ GeV, but what happened before that is not known. Suppose that, for some reason, the dependence of a on t before baryogenesis was such that the integral in (98) is very large and the factor aH increases with time rather than decreases. Then the problem of initial conditions could be solved naturally.

To get a better grasp of the idea, let us see how it works explicitly. Suppose that the Universe was dominated by the vacuum energy at some point before the baryogenesis, BBN, and photon decoupling. As we found earlier, a vacuum-energy-dominated Universe expands as

$$a(t) = a_0 \exp\left[H(t - t_0)\right] \tag{103}$$

with $H = \sqrt{8\pi G\rho_\Lambda/3} = \text{const}$, where ρ_Λ is the vacuum energy density dominating the energy budget of the Universe during the period of inflation. Then, at $t = t_1$ the vacuum-energy dominated epoch ends with a transition to the radiation-domination epoch. After that, the scale factor continues to evolve as

$$a(t) = a_0 \exp\left[H(t_1 - t_0)\right] \left(\frac{t}{t_1}\right)^{1/2} . \tag{104}$$

Hence, using (98) we can calculate the horizon size at photon decoupling as

$$\begin{aligned} \ell_H(t_d) &= \int_{t_0}^{t_1} dt' \exp\left[H(t-t')\right] + \int_{t_1}^{t_d} dt' \left(\frac{t}{t'}\right)^{1/2} \\ &= \frac{1}{H} \exp\left[H(t_1 - t_0)\right] + 2(t_d - t_1) \\ &\sim \frac{1}{H} \exp\left[H(t_1 - t_0)\right] , \end{aligned} \tag{105}$$

where in the last step we have assumed that the period of inflation was sufficiently long that $H(t_1 - t_0) \gg 1$ and $H^{-1} \exp\left[H(t_1 - t_0)\right] \gg 2(t_d - t_1)$. If, for example, $\rho_\Lambda \sim \left(10^{15}\ \text{GeV}\right)^4$ (GUT scale), the horizon problem is solved if the particle horizon at the time of photon decoupling as seen today $\ell_H(t_d)a_0/a_d$ is at least as large as the present Hubble radius H_0^{-1}. The latter requirement gives a lower bound on the number of e-foldings during inflation $H(t_1 - t_0) \gtrsim 65$. The flatness problem is also automatically solved since

$$\left.\frac{\kappa}{a^2H^2}\right|_{t=t_0} \propto \kappa \exp\left[-2H(t_1 - t_0)\right] \tag{106}$$

becomes exponentially suppressed if $H(t_1 - t_0) \gg 1$. The time-dependence of κ/a^2H^2 is, of course, more complicated than displayed above, but the point is that as long as $H(t_1 - t_0) \gg 1$ the exponential suppression will most likely be the dominant factor.

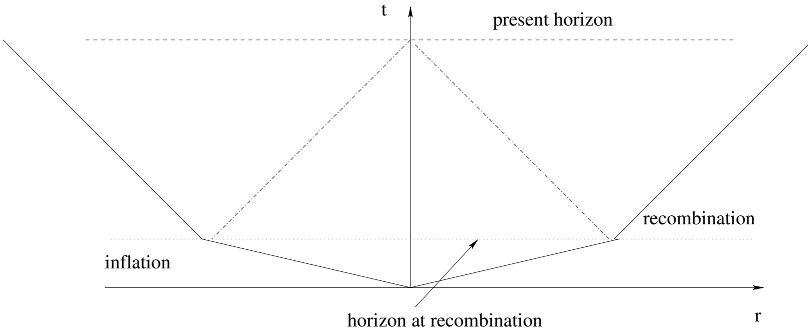


Fig. 9: An observer at point $r = 0$ sees just one particle horizon corresponding to photon decoupling.

14.3 Chaotic inflation

There are many different particle-physics models of inflation. Most of them are associated with the dynamics of single or multiple scalar fields. We refer to a comprehensive review [16] and describe in some detail just one possibility which is called “chaotic inflation”.

Consider a theory of a single free scalar field in a curved background with an action

$$S = \int d^4x\sqrt{-g}\left(\frac{1}{2}g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - U(\phi)\right) , \tag{107}$$

where the potential is

$$U(\phi) = \frac{1}{2}m^2\phi^2 . \tag{108}$$

We will assume that $m^2 \ll M_{Pl}^2$.

We do not know how to describe the state of the Universe at the Planck scale, since the classical theory of gravity is not applicable there. Nevertheless, it is natural to assume that at Planck time there were fluctuations in the scalar field with energy density

$$\epsilon \sim \frac{1}{2}\dot{\phi}^2 + \frac{1}{2}(\nabla\phi)^2 + U(\phi) \sim M_{Pl}^4 . \tag{109}$$

Suppose that there is a sufficiently large region of space where fluctuations of potential energy dominate, i.e.

$$U(\phi) \sim M_{Pl}^4 \gg (\nabla\phi)^2 \text{ and } \dot{\phi}^2 . \tag{110}$$

In these regions, the value of ϕ is much larger than the Planck scale,

$$\phi \simeq \frac{M_{Pl}^2}{m} \gg M_{Pl} , \tag{111}$$

and the scalar field is nearly homogeneous so that the equation representing its evolution is simply

$$\ddot{\phi} + 3H\dot{\phi} + \frac{dU(\phi)}{d\phi} = 0 , \tag{112}$$

where

$$H^2 = \frac{8\pi G}{3}\left(\frac{\dot{\phi}^2}{2} + U(\phi)\right) . \tag{113}$$

This looks like an equation of motion of a non-relativistic particle with unit mass in the potential $U(\phi)$ with a friction term that depends on the position and velocity of the particle, see Fig. 10. For large values of ϕ the regime is overdamped, with $H \gg \ddot{\phi}/\dot{\phi}$, where

$$H^2 \simeq \frac{4\pi m^2\phi^2}{3M_{Pl}^2} . \tag{114}$$

Thus, eq. (112) has the form

$$\frac{\sqrt{12\pi}m\phi\dot{\phi}}{M_{Pl}} + m^2\phi = 0 \tag{115}$$

and has a solution

$$\phi \simeq \phi_0 - \frac{mM_{Pl}t}{\sqrt{12\pi}} \simeq \phi_0\left(1 - O\left(\frac{m^2t}{M_{Pl}}\right)\right) . \tag{116}$$

The “slow-roll-down" approximation breaks down at $\dot{\phi}^2 \sim V(\phi)$, where $\phi \simeq M_{Pl}$, $t \simeq M_{Pl}/m^2$.

Before this time the Universe expands exponentially and the scale factor changes by

$$\exp\left(Ht\right) \simeq \exp\left(\frac{M_{Pl}^2}{m^2}\right) \gg 1\;. \tag{117}$$

During exponential expansion, the non-homogeneities are red-shifted away and the Universe becomes practically uniform at cosmological distances.

After time $t \simeq M_{Pl}/m^2$ inflaton oscillates near the origin, transferring its energy to other particles. This process is usually called reheating. For $m \ll M_{Pl}$ this simple model of scalar field solves horizon, homogeneity and flatness problems. It does not survive, though, a more elaborated comparison with the data concerning the spectrum of primordial density perturbations, which eventually lead to structure formation in the Universe, see below. However, replacing a free field theory with a more complicated "inflationary" potential allows us to address all the issues.

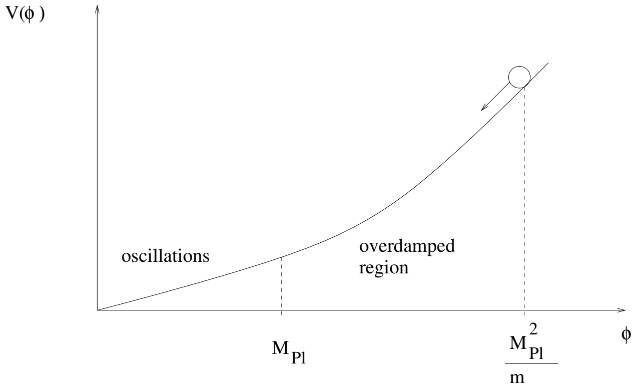


Fig. 10: Inflaton dynamics as the motion of a particle in potential $V(\phi)$ with friction term depending on ϕ.

14.4 Density perturbations from inflation

The paradigm of inflation was originally motivated by solutions to the horizon and flatness problems. It was realised soon after that, the quantum aspect of the same idea could provide the appropriate initial conditions for the primordial density perturbation.

Qualitatively, inflation seeds the density perturbations in the Universe in the following way. Quantum fluctuations $\delta\phi$ in the inflaton field result in adiabatic density perturbation $\delta\rho(\mathbf{x})$ on top of the homogeneous background density ρ. Existing initially at small distances, perturbations are inflated to macroscopic scales due to the exponential expansion of the Universe at the inflationary stage. By definition, the average

$$\left\langle\frac{\delta\rho(\mathbf{x})}{\rho}\right\rangle = 0\,, \tag{118}$$

but

$$\left\langle\frac{\delta\rho(\mathbf{x})}{\rho}\frac{\delta\rho(\mathbf{y})}{\rho}\right\rangle = f(|\mathbf{x}\text{-}\mathbf{y}|) \neq 0\,, \tag{119}$$

Note that f depends only on the $|\mathbf{x} - \mathbf{y}|$ as a consequence of isotropy and homogeneity. Due to inflation, for $\mathbf{x}$ and $\mathbf{y}$ located within a causal domain $|\mathbf{x} - \mathbf{y}| \lesssim \ell_H(t_0)$, the correlation function should be close to constant

$$\left\langle \frac{\delta\rho(\mathbf{x})}{\rho} \frac{\delta\rho(\mathbf{y})}{\rho} \right\rangle \simeq \text{const} \,. \tag{120}$$

As it is impossible to have a perfectly constant Hubble parameter during inflation, there is a slight tilt in the correlation function (the Fourier transform of it defines the so-called power spectrum of primordial fluctuations)

$$f(\mathbf{x}) \propto |\mathbf{x}|^{1-n_s} \,, \tag{121}$$

where $n_s \neq 0$ is the so-called spectral index of the scalar perturbations. A scale-invariant spectrum with $n_s = 0$ is called the Harrison-Zeldovich power spectrum.

In addition to scalar perturbations $\delta\rho/\rho$, inflation also predicts tensor perturbations. It is common to measure their amplitude relative to the size of the scalar perturbation in terms of the tensor-to-scalar ratio

$$r = \frac{\rho_2}{\rho_0} \,, \tag{122}$$

where ρ_2 is the energy of spin-2 fluctuations (associated with the gravitational waves created during inflation) and the energy in the scalar fluctuations. These two parameters, n_s and r depend on the model of inflation.

To explain the CMB temperature and polarisation fluctuations, inflation gives as input the form of the power spectrum characterised by n_s and tensor to scalar ratio r. The initial density perturbations get modified by various "ordinary physics" effects, e.g. photon scattering, gravitational lensing, plasma waves, etc that we know how to incorporate. Once these effects are accounted for, one can obtain, for instance, a prediction for the present form of temperature fluctuation $\left\langle \frac{\delta T(\mathbf{x})}{T} \frac{\delta T(\mathbf{y})}{T} \right\rangle$ which we can compare with CMB measurements. The predicted form depends on many parameters, e.g. n_s, r, Ω_B, Ω_ν, etc, enabling themselves to be determined by measuring the CMB temperature and polarisation anisotropies. See Figure at slide 18 of Lecture 3.

The whole picture, called ΛCDM (for Lambda-Cold Dark Matter cosmology) works remarkably well, with only a few fitting parameters describing a huge amount of cosmological data (but several tensions exist, for instance, the Hubble rate measurements at small and large redshifts give different results).

14.5 Particle physics models for inflation

What kind of particle inflated the Universe? Is this a new particle or a part of the Standard Model? The answer is: “We do not know”. There is a huge variety of inflationary models based on different principles and ideas, for a review see [16]. The mass of the inflaton may vary from hundreds of MeV to, say, 10^{10} GeV. The minimal possibility is that the Higgs field of the Standard Model inflated the Universe is in perfect agreement with the data on the spectral tilt n_s and tensor to scalar ratio r (see discussion on slides 19-22 of Lecture 3). To single out the inflationary model we need a theory input in the form of the complete theory of gravity and the SM, together with an even more precise determination of the cosmological parameters allowing us to discriminate the different models.

15 Conclusions

The remarkable progress in particle physics and cosmology resulted in the creation of two standard models – the Standard Model in particle physics and ΛCDM in cosmology, both consistent with the majority of observations in respective domains of physics. There is a clash between these two theories, though. Indeed, only a small part $\sim 5\%$ of the energy density of the Universe can be described by particles of the Standard Model - baryons. Even this number, associated with the baryon asymmetry of the Universe, cannot be explained by the physics of the SM. Another $\simeq 70\%$ can be attributed to the cosmological constant or "Dark Energy" with the equation of state close to that of the cosmological constant. The remaining $\simeq 25\%$ represents dark matter, which cannot be the SM particles. The SM, where neutrinos are massless and neutrino flavours are conserved, is also challenged by neutrino physics.

There are many theoretical ideas in particle physics and cosmology which address the drawbacks of the SM. The structure of the Universe at large scales and density perturbations are neatly explained by inflation. The problems of neutrino masses, baryon asymmetry of the Universe and Dark Matter can be solved in different extensions of the SM. Hopefully, new observations in cosmology and new experiments in particle physics will narrow the search for the theory which provides a better description of Nature.

References

[1] S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*. John Wiley and Sons, New York, 1972.

[2] Y. B. Zeldovich and I. D. Novikov, *Relativistic Astrophysics. vol. 2. The Structure and Evolution of the Universe*. 1983.

[3] E. W. Kolb and M. S. Turner, *The Early Universe*, vol. 69. 1990.

[4] V. A. Rubakov and D. S. Gorbunov, *Introduction to the Theory of the Early Universe: Hot big bang theory*. World Scientific, Singapore, 2017.

[5] D. S. Gorbunov and V. A. Rubakov, *Introduction to the theory of the early universe: Cosmological perturbations and inflationary theory*. 2011.

[6] V. Mukhanov, *Physical Foundations of Cosmology*. Cambridge University Press, Oxford, 2005.

[7] A. G. Cohen, D. B. Kaplan, and A. E. Nelson, "Progress in electroweak baryogenesis," *Ann. Rev. Nucl. Part. Sci.* **43** (1993) 27–70, `arXiv:hep-ph/9302210`.

[8] S. Davidson, E. Nardi, and Y. Nir, "Leptogenesis," *Phys. Rept.* **466** (2008) 105–177, `arXiv:0802.2962 [hep-ph]`.

[9] D. E. Morrissey and M. J. Ramsey-Musolf, "Electroweak baryogenesis," *New J. Phys.* **14** (2012) 125003, `arXiv:1206.2942 [hep-ph]`.

[10] L. Canetti, M. Drewes, and M. Shaposhnikov, "Matter and Antimatter in the Universe," *New J. Phys.* **14** (2012) 095012, `arXiv:1204.4186 [hep-ph]`.

[11] A. D. Dolgov, "NonGUT baryogenesis," *Phys. Rept.* **222** (1992) 309–386.

[12] A. Riotto and M. Trodden, "Recent progress in baryogenesis," *Ann. Rev. Nucl. Part. Sci.* **49** (1999) 35–75, `arXiv:hep-ph/9901362`.

[13] R. Allahverdi and A. Mazumdar, “A mini review on Affleck-Dine baryogenesis,” *New J. Phys.* **14** (2012) 125013.

[14] A. Boyarsky, O. Ruchayskiy, and M. Shaposhnikov, “The Role of sterile neutrinos in cosmology and astrophysics,” *Ann. Rev. Nucl. Part. Sci.* **59** (2009) 191–214, `arXiv:0901.0011 [hep-ph]`.

[15] A. Boyarsky, M. Drewes, T. Lasserre, S. Mertens, and O. Ruchayskiy, “Sterile neutrino Dark Matter,” *Prog. Part. Nucl. Phys.* **104** (2019) 1–45, `arXiv:1807.07938 [hep-ph]`.

[16] J. Martin, C. Ringeval, and V. Vennin, “Encyclopædia Inflationaris,” *Phys. Dark Univ.* **5-6** (2014) 75–235, `arXiv:1303.3787 [astro-ph.CO]`.

Practical statistics excerpts from 2023 European School of High Energy Physics

Troels C. Petersen[a]

[a]Niels Bohr Institute, Copenhagen, Denmark

These lecture notes provide an overview of fundamental statistical concepts used in high-energy physics research. The discussion begins with the philosophy of statistics, emphasizing its role in analyzing experimental data, correcting biases, and ensuring precision. Key estimators such as mean, standard deviation, skewness, and kurtosis, which help summarize datasets, are introduced. Various probability density functions (PDFs), including Binomial, Poisson, and Gaussian distributions, are explored to demonstrate their relevance in data modeling. Methods for analyzing data, such as maximum likelihood estimation and ChiSquare tests, are presented, offering techniques for extracting meaningful insights from observations. The notes also cover hypothesis testing, explaining concepts like false positive rates, p-values, and significance levels for evaluating scientific claims. Practical applications include setting observational limits and improving statistical methodology. By applying these techniques, researchers can ensure rigorous data analysis, enabling reliable conclusions and impactful discoveries in physics.

This chapter should be cited as: Practical statistics excerpts from 2023 European School of High Energy Physics, Troels C. Petersen, DOI: 10.23730/CYRSP-2025-009.103,
in: Proceedings of the 2023 European School of High-Energy Physics,
CERN Yellow Reports: School Proceedings, CERN-2025-009, DOI: 10.23730/CYRSP-2025-009, p.103.

1 Philosophy of statistics

When confronted with hard-won data, one of course wants to analyse it in the most optimal way. However, this is often a bit harder than one might think.

> "The art of drawing conclusions from experiments and observations consists in evaluating probabilities and in estimating whether they are sufficiently great or numerous enough to constitute proofs. This kind of calculation is more complicated and more difficult than it is commonly thought to be."
>
> [Antoine Lavoisier, French chemist 1743-1794]

1.1 Why Statistics?

We collect data in order to see trends and compare it to our expectations (various theories in physics). However, experiments are inherently non-deterministic due to both quantum and chaos effects. Even without these, experiments are limited in precision by cost, time, etc. We are thus limited to finite samples with limited resolution that typically need to be corrected for all sorts of nuisance effects in order to be both accurate (i.e. non-biased) and precise (i.e. with a small uncertainty). Statistics is the tool for this process, and will—along with domain knowledge—be needed for extracting good estimates with correct uncertainties.
And given the hard work that lies behind planning, building, and running large detectors in order to give birth to cutting edge data, it would be near criminal to analyse it with anything but the most powerful data analysis methods available.

A common misconception is that statistics provides a straight path forward in all situations. This is not so. There are of course great examples of clear cut cases, but more often it is wise to follow the insights of John Tukey (US statistician, 1915-2000) who argued "the need for statisticians to reject the role of 'guardian of proven truth', and to resist attempts to provide once-for-all solutions and tidy over-unifications of the subject".

1.2 Why Uncertainties?

A single number without any indication of the size of its uncertainty is useless. To see this, imagine that you had measured the speed of gravity to be 2.91×10^8 m/s. Perhaps surprisingly, such a measurement would tell you…nothing! For depending on the size of the uncertainty, you might reach three very different conclusions:

$(2.91 \pm 7.36) \times 10^8$ **m/s.** In this case, the speed of gravity could be pretty much anything, also far exceeding the speed of light or even negative.[1]

$(2.91 \pm 0.07) \times 10^8$ **m/s.** This result is consistent with the speed of light and not much else.

$(2.908 \pm 0.007) \times 10^8$ **m/s.** Here, the small uncertainty (high precision) of the measurement shows that it is NOT exactly the speed of light, and hence a new discovery of some phenomena that slows down the speed of gravity by about 2.5%.

One of the main goals of doing good experiments and, hence. physics research is to minimise the uncertainties on the results. One of the main goals of data analysis and, hence, statistics is to get the uncertainties right! Obtaining credible uncertainties is hard work bordering on an art form. While statistical methods yield known and well understood uncertainties, it is your job as an experimenter and expert to check that all sources of uncertainty are accounted for. In the end, the target is to estimate the values, distributions, and principles behind the origin of the data, thereby "luring" the truth out of nature (in science), the body (medicine), and the markets (economy).

2 The basics of statistics

The simplest data is a series of N independent observations or measurements of the same quantity, $x_1, \ldots, x_n$. It could consist of anything from heights of people to invariant masses of decayed elementary particles. The first thing we want to do is inspect, describe, and summarize the data.

2.1 Printing and plotting data

The simplest way that we can "see" the data visually is by printing a few values and plotting the data in a histogram. When reading a new data file, it is worthwhile to print the first 10–20 entries along with the total number of entries, just to get a crude impression of the data types and sizes. Make sure that you know if the data is integer, discrete, rounded, continuous, or contains Not-a-Number (NaN) or outlier values.

The obvious next step is plotting the data. For a single data series (i.e. 1D data) a histogram is the obvious choice. Always make sure that you control the histogram range and number of bins, and do so cleverly! Algorithms (ROOT or MatPlotLib) are notoriously poor at doing this, as they do not have any sense of what is important. You should for integers make sure, that the bin width is an integer, and that it is shifted by a half to ensure that the middle of the bin matches the average of the bin range (Example: For the number of hits in a certain detector with up to 20 hits, choose the range to be $[-0.5, 20.5]$ with 21 bins). The histograms shows the data distribution, and "transforms" the (potentially very long) list into x-values (middle of the bins), y-values (number of entries in each bin), and σ_y-values (uncertainty on the y-values, approximate by the square root of the y-values themselves), which in turn can be fitted (see Fitting Data).

[1] Given the size of the uncertainty, one would also report the result as $(3 \pm 9) \times 10^8$ m/s.

More advanced is plotting pairs of data points (y-values vs. x-values, times series, etc.) producing either a graph (ordered data) to see trends or a scatter plot (unordered data) to see correlations.

2.2 Estimators

While printing and plotting data is imperative, statistics is about quantification: We want to assign numerical values to data, that in simple terms describe the data beyond the (important) number of data points. For this, we use estimators. These are functions of the data, which yield a single value output (the estimate). The typical notation for an estimator is the quantity it is trying to estimate (e.g. μ) with a hat above (as in $\hat{\mu}$).

In order to describe the distribution that has given rise to a dataset, we first of all want to convey the typical value of x. This can be done in many different ways, most often based on the (arithmetic) mean defined as:

$$\hat{\mu}(x) = \frac{1}{N} \sum_i x_i \; . \tag{1}$$

While the formula for the mean is of course well known to all, it is worth noting that this formula arises from the principle of maximum likelihood (as most basic formulae do in statistics) and also while it is a great estimator, it is not perfect for all purposes, as it is not very robust to far outliers (imagine values around 0–1 with a single way-off value of 1000). There are many alternatives: Median, Mode, Geometric Mean, Harmonic Mean, and Truncated Mean. The median and truncated mean are both robust to outliers, while the others hold other virtues. Yet, the usual (arithmetic) remains the most commonly used.

In addition to a typical value, the next thing to consider would be the typical variation of the values, thus also giving an idea of the range of values. This is described by the variance V, or rather the square root of it, called the *Standard Deviation* (SD or Std.),[2] defined as:

$$\widehat{Std}(x)^2 = V(x) = \frac{1}{N} \sum_i (x_i - \mu)^2 \; . \tag{2}$$

Given a dataset, we don't know the true mean μ. Naively, we could calculate the sample Std. using the estimated mean from above instead of the true mean. However, then we have used some information (one degree of freedom) for estimating this mean, and hence one should correct the formula (known as Bessel's correction) for the Std. as follows:

$$\widehat{Std}(x)^2 = V(x) = \frac{1}{N-1} \sum_i (x_i - \mu)^2 \; . \tag{3}$$

This is an unbiased estimator of the Std., as can be proven mathematically or shown by example using a simple simulation (e.g. take the Std. of N=3 random unit Gaussian numbers many times, and consider the distribution mean of the Std. values when including the correction or not).

Given repeated measurements of the same quantity, the Std. describes the typical (Gaussian) uncertainty (denoted $\sigma(x)$) on *each measurement*. However, as the precision of the mean improves with the square

[2] The reason why both the variance and the square root of it (the Std.) has a specific name in statistics is likely, that the variance has theoretical importance and is used in many calculations, while the Std. has the same unit as the values we consider and is therefore often the number that we seek, understand easiest, and therefore quote.

root of the number of measurements, the uncertainty *on the estimated mean* ($\sigma(\hat{\mu}_x)$) is given as:

$$\sigma(\hat{\mu}_x) = \sigma(x)/\sqrt{N}\ . \tag{4}$$

Make absolutely sure that you understand the difference between the uncertainty on a single measurement (σ_x) and the uncertainty on an estimated mean ($\sigma(\hat{\mu}_x)$), as this is a very common (and grave) mistake to make!

EXAMPLE:

In 1797–98 Henry Cavendish famously made 29 independent measurements of the average density (and thus the total mass) of the Earth:
5.5, 5.61, 5.88, 5.07, 5.26, 5.55, 5.36, 5.29, 5.58, 5.65, 5.57, 5.53, 5.62, 5.29, 5.44, 5.34, 5.79, 5.1, 5.27, 5.39, 5.42, 5.47, 5.63, 5.34, 5.46, 5.3, 5.75, 5.68, and 5.85.
After having inspected the values and plotted them in a histogram (e.g. 10 bins in the range [5.0, 6.0], as the extreme values are 5.1 and 5.85), we calculate the mean to be 5.448 and the Std. to be 0.333. This in turn gives an uncertainty on the mean of $0.33/\sqrt{29} = 0.06$, and we would thus give the result as 5.448 ± 0.06g/cm^3.
Note: It took more than a century to improve on this great measurement, which was within 1% of the modern day value.

The Std. has many names, among others the Root Mean Square Error (RMSE, where the name contains the recipe for its calculation), and simply the width of a distribution. For a Gaussian distribution, which is given by:

$$G(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{5}$$

the Std. coincides with the σ in the formula.
The uncertainty on the Std. (“the uncertainty on the uncertainty”) is rarely considered, as quantifying the uncertainty alone is typically sufficient. However, for adequate statistics, it can be estimated as:

$$\sigma(\text{Std.}) = \text{Std.}/\sqrt{2(n-1)} \tag{6}$$

The mean and width of a distribution can be seen as the 1st and 2nd (central) moment of data. Naturally, higher moments exists, named the Skewness (3rd) and Kurtosis (4th), defined as:

$$\text{Skewness} = \sum_i (x_i - \mu)^3/\sigma^3 \tag{7}$$

$$\text{Kurtosis} = \sum_i (x_i - \mu)^4/\sigma^4 - 3 \tag{8}$$

The Skewness measures the asymmetry of a distribution, while the Kurtosis quantifies how long the tails of a distribution are. Both are 0 for the Gaussian distribution, which (of course) holds a special place in statistics.
Taking a step back, and thinking about estimators, these are meant to be good prediction for a true value

that we don't know. But what do we mean by "good"? This is defined by three criteria:

Consistency. The estimator must converge towards the true value for large statistics.

Asymptotic normality. The estimator should obtain a normal distribution around the true value for large statistics.

Efficiency. The estimator must have the minimal possible error, defined by the so-called Rao–Cramer-bound.

While estimators can provide key values that quantify distributions, this does not necessarily give any insight into the type and hence origin of the distributions at hand. This requires that we recognize certain typical Probability Distributions Functions (PDFs), which are the result of some (typically simple) underlying principles.

2.3 Covariance and correlation

When given two variables x and y, one can estimate the covariance between them V_{xy} as follows:

$$V_{xy} = \frac{1}{N-1}\sum_{i}(x_i - \mu_x)(y_i - \mu_y)\ . \tag{9}$$

The covariance quantifies to which degree x and y are *linearly* correlated. If y is high (above its mean) when x is high (above its mean) and similarly for low values, then the covariance obviously comes out positive. This indicates a positive correlation. Conversely, if y is high when x is low and vice versa, then V_{xy} becomes negative, indicating a negative correlation.

Extending Eq. (9) to include all combination of variables, this produces a matrix of variances, the so-called covariance matrix, which is symmetric, and where the diagonal carries the variances of each variable. The correlation matrix plays a central role in statistics. It encapsulates to what degree the different variables are related to one another, both for measurements (as Eq. (9) describes), but also for e.g. fit parameters.

While the numerical value of the covariance can be hard to interpret, the correlation coefficient ρ_{xy} defined below is much easier to understand, as it only takes values in the range $[-1,1]$:

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}\ . \tag{10}$$

It is important to notice that in the above only linear correlations can be determined. Non-linear correlations are not included. Thus, if $\rho \neq 0$ one can conclude that there is a correlation. However, if $\rho = 0$, then one can *not* be sure that there are no correlations. If $\rho = \pm 1$ then x and y are completely correlated.

Correlation can also be measured based on rank (i.e. position when sorted) rather than value. One example is Spearman's rank correlation, which follows the same formula as for ρ, but where the values (x,y) in the covariance formula have been substituted with their ranks ($R(x)$, $R(y)$), that is the (integer) position in a list sorted by value. The philosophy behind rank correlation is that it measures the degree of bijectivity, i.e. if there is a monotonic relation between values of x and y, even if it is not linear.

There exists more complex measures of correlation, which also includes non-linear relations through the amount of information that x and y share. Some of these measures are Distance Correlation, Correlation Ratio, and Mutual Information (based on the Kullback–Leibler divergence—beyond the scope of these notes).

3 Probability Density Functions

A Probability Density Function (PDF) is a function, $f(x)$, that describes the probabilities of specific outcomes. They can be discrete (throw of a die: $X \in [1, 2, 3, 4, 5, 6]$) or continuous (random number from computer: $x \in [0, 1]$). The value of a PDF at a specific point, x_0, can be interpreted as a relative likelihood for a random value from $f(x)$ to take on the value x_0. Thus, $f(x_0)dx$ is the probability of x falling in the infinitesimal interval $[x_0, x_0 + dx]$.

In order for a function to be a PDF, it must never produce negative values (no negative probabilities) and must have a unit integral, so that probability is conserved (i.e. the probability of any outcome is one): $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)dx = 1$. For discrete PDF distributions these formulae can be considered as sums.

The mean and variance of a PDF are calculated much like their corresponding estimators (note: These are the true values, not estimates from a sample, and hence there are no hats):

$$\mu = \int_{-\infty}^{\infty} x f(x)dx \; , \tag{11}$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \; . \tag{12}$$

Since PDFs “only” provides values proportional to probabilities, the integral of PDFs is often of interest. Specifically, the Cumulative Distribution Function (CDF) and Survival Function (SF) are defined as:

$$CDF_{F(x)} = \int_{-\infty}^{x} f(t)dt \; , \tag{13}$$

$$SF_{f(x)} = \int_{x}^{\infty} f(t)dt \; . \tag{14}$$

Obviously, CDF(x) + SF(x) = 1, and both are just a simple way of writing the specific integrals in a simple way. We will return to the SF, when we get to the ChiSquare method.

3.1 The Central Limit Theorem and the Gaussian distribution

Imagine that you add some random numbers and repeat the process (in the same way) many times. What distribution of sums should you expect? While both the question and the answer may seem a bit arbitrary, the answer and the implications are rather surprising: The distribution of sum resembles a Gaussian distribution! And that is why uncertainties tends to be Gaussian! More specifically, it can be proven that “the sum of N independent continuous random variables x_i with means μ_i and variances σ_i^2 becomes a Gaussian random variable with mean $\mu = \sum_i \mu_i$ and variance $\sigma^2 = \sum_i \sigma_i^2$ in the limit that N approaches infinity.”

This is called the Central Limit Theorem (CLT), and holds a prominent place in probability theory, as it is much of the foundation behind our thinking and working with random variations, uncertainties, and the Gaussian as the “unit distribution” of statistics.

You might think, that adding random numbers of various unknown distributions is a special case, but it turns out, that most continuous measurements are exactly a result of such a process: You start with a specific “true” value (e.g. energy of a photon from a specific decay), which is perturbed through interactions before reaching your detector, state of your detector at the time of arrival, random processes in relation to the photon interactions in the detector, distortions from the read-out electronics, etc. to end at a value around the true value, but not quite. The CLT is the reason why you obtain a distribution resembling a Gaussian, when repeating a measurement many times. And if it is not Gaussian, but exhibits a clear structure, then typically some effect is at play, which you will want to find, understand, and correct for.
While the Gaussian distribution holds a central place in statistics, there are several other fundamental distributions, which are important.

3.2 Discrete PDFs

3.2.1 Binomial distribution

Consider a random process (e.g. roll of a die) with probability of success (e.g. rolling a 6) p repeated independently N times. The distribution of number of successes n (denoted $P(X = n)$) will then follow a Binomial distribution:

$$f(n, N, p) = P(X = n) = \left(\frac{N!}{n!(N-n)!}\right) p^n (1-p)^{N-n} \tag{15}$$

The formula is relatively straightforward to understand. The n successes have probability p^n while the $N-n$ failures have probability $(1-p)^{N-n}$ and the number of ways these can be obtained is the binomial coefficient, defined as:

$$\binom{N}{n} = \left(\frac{N!}{n!(N-n)!}\right) \tag{16}$$

If you sum all the terms in the Binomial distribution, you should see that you get 1 no matter the values of the two parameters p and N, so the PDF is normalised, as it should be.
The mean and variance of the Binomial are $\mu = Np$ and $\sigma^2 = Np(1-p)$, respectively. As the Binomial is often used in relation to fractions $f = n/N$ (e.g. efficiencies), the formula for the uncertainty on a fraction is useful to remember: $\sigma(f) = \sigma(n)/N = \sqrt{Np(1-p)}/N = \sqrt{p(1-p)/N}$.
The requirements for being binomially distributed are a fixed number of trials (N) that are independent and with only two outcomes of constant probability. Thus, e.g. the cards of a poker hand or rolling a die until a 6 appears are not binomially distributed. If there are more than two outcomes, the formula expands to become the multinomial distribution.

3.2.2 Poisson distribution

Sometimes, one does not know p and N separately for a process, but only the rate of success (whatever that defines), given as $\lambda = Np$, and then the Binomial distribution can not be used. But if N is large (which also makes the calculation of the binomial coefficients challenging) and p is small, then the Binomial distribution of number of successes approaches the Poisson distributed (by Stirling’s formula):

$$f(n, \lambda) = P(X = n) = \frac{\lambda^n e^{-\lambda}}{n!} \tag{17}$$

Notice how the Poisson distribution only has one parameter, λ, and that both the mean and the variance is λ. The latter is of tremendous importance to remember (i.e. commit to memory now!), as this means that the uncertainty on a (Poisson) number is the square root of that number. You may think that being a Poisson case is rare, but in fact most numbers are:

Number of entries in a single bin in a histogram. This is why we assign bin uncertainties as the square root of the number of entries in a bin. It is of course an approximation that these uncertainties are Gaussian, but if $\lambda > 20$ then it is a very good approximation.

Number of people killed in traffic a year. Many venture into traffic many times in a year, and the probability each time is (fortunately) very small. The probability is not constant, but the sum of numbers from two Poisson distributions (λ_a and λ_b) is also Poisson distributed (λ = λ_a + λ_b).

Number of die hard left-wings voting for the conservatives. Just as another example, where p is very likely small, as is required.

In fact, both the Binomial and the Poisson distributions tend to take the shape of a (discrete) Gaussian distribution for large values of Np, as do many other continuous PDFs in other limits corresponding to high statistics.

EXAMPLE:

The square root uncertainty on a counting number has a significant and growing impact. Take the number of people killed in traffic a year, which in 2022 was 155 in Denmark. The statistical uncertainty is $\sqrt{155}$ = 12.4 or about 8%. Thus, in trying to improve traffic safety, only effects larger than this can be seen with a single year's data.

Consider then the situation for commercial flights ending in deaths. Perhaps there are say four each year, that is with a 50% uncertainty! How will you ever know if new initiatives to improve aviation safety have any impact? Wait 50 years?

The answer was to change the legislation completely! By now, everyone is obligated to report near-misses, faulty equipment, etc., at the risk of getting fired if NOT reporting. There are 10000s of such cases, giving plenty of statistics to monitor developments in detail. Such changes in legislation is becoming more and more common (e.g. entering hospitals), all because of the Poisson distribution.

3.3 Continuous PDFs

3.3.1 Uniform

The uniform distribution is perhaps best known for being the random numbers that computers provide, but it can of course be made more general between a and b:

$$f(a,b) = \frac{1}{|b-a|} \text{ for } x \in [a,b], \text{ else } 0 \tag{18}$$

The unit uniform PDF has mean 0.5 and variance 1/12. You can calculate these values from Eqs. (12), which is a good exercise to check that you have understood the concept. Given that uniform numbers are the typical computer output, this is also those to be transformed into other distributions.

3.3.2 Exponential distribution

The exponential distribution is single parameter PDF given by:

$$f(\tau) = \frac{1}{\tau} \exp(-t/\tau) \ \text{ for } x \in [0, \infty] \ . \tag{19}$$

The distribution is in particle physics typically used for lifetime measurements and background modelling.

3.3.3 ChiSquare distribution

The ChiSquare distribution is in fact a family of distributions defined by the number of degrees of freedom (N_{DOF} or ν). It is essentially the distribution obtained from adding ν unit Gaussian numbers squared.

$$f(\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} \exp(-x/2) \ \text{ for } x \in [0, \infty] \ . \tag{20}$$

The distribution is interesting, since it reflects the expected χ^2 values from e.g. a comparison or fit with the corresponding N_{DOF}. This is central for testing goodness-of-fit in χ^2-fits and in hypothesis testing. Note that the mean value is ν.

3.3.4 Student's t distribution

The Student's t distribution is the curious result of beer brewing and industrial secrecy, and it is an example of Stigler's Law of Eponymy. Like the ChiSquare distribution, it only has ν as a parameter:

$$f(\nu) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)} \left(1 + t^2/\nu\right)^{-(\nu+1)/2} \ \text{ for } t \in [-\infty, \infty] \ . \tag{21}$$

The distribution is a generalisation of the Gaussian distribution, which it approximates for large values of ν, while it is the "infinite tailed" Cauchy distribution for $\nu = 1$. It is used when doing hypothesis testing with small samples, where the mean and variance are not well known but estimated. Once the statistics gets significant (typically above 10), the Gaussian distribution becomes a good approximation for the hypothesis testing.

3.3.5 Other distributions

There is a wealth of different PDF distributions, many of which are used for very special situations. For this reason note that polynomials are typically not included as PDFs, as these can take on negative values and are usually not normalised. These deficiencies can be rectified considering special classes of polynomials, but natural phenomena rarely follow (higher order) polynomial distributions. Polynomials are therefore rarely used for modelling histogram data, but rather for fitting points with uncertainties. Of course, occasionally one resorts to using a polynomial modelling of the background distribution, especially when the shape is complicated, statistics is high, and the range of interest is finite (e.g. $H \to \gamma\gamma$).

3.4 The philosophy of PDFs

While the basic PDFs often serve well as the building blocks for models, they certainly have their limitations, famously summarised as:

> "Essentially, all models are wrong, but some are useful"
> [George E. P. Box, British Statistician 1919–2013]

The point is, that real data is never ideal and hardly ever follows these simple PDFs perfectly. Especially, when considering a lot of data, models often tend to have a hard time mimicking the data perfectly, as the high statistics and corresponding small uncertainties makes all the minuscule variations and unknown effects stand out.

This is expected, and the reader should ensure awareness and understanding of the imperfect modelling, rather than introducing new ad hoc fitting parameters. No need to chase mice, when there are tigers around!

4 Fitting data

Given data, we typically want to extract information from it. Going beyond the estimators introduced in section 2.2, this would typically consist of fitting the data with a functional form.

Data typically consist of points (x, y) with an uncertainty on y, $\sigma(y)$ (and potentially also on x, $\sigma(x)$), where the uncertainty is considered Gaussian. Alternatively, the data is a series of measurements, which can be put into a histogram, typically in 1 dimension (1D), but potentially in 2D or even 3D. We will in the following only consider fitting in 1D, though the arguments and methods extend to multiple dimensions. However, be warned that the complexity of fits grows fast with increasing dimensionality, due to possible correlations, large number of fit parameters, and an explosion in the number of bins.

Testing to what extent the fit model—and more generally any model, from fitting or not—actually matches the data is a core theme in statistics. For now, we will start by considering how to perform a fit, and how to extract results and uncertainties from it. Towards the end, we will start asking to what extent the model is reasonable, naturally leading to the next chapter on hypothesis testing.

4.1 Principle of Maximum Likelihood

Fitting data means obtaining the fit parameter values such that the function "best matches" the data. Obviously, we need to define what we mean by "best matches". Here the Principle of Maximum Likelihood (PML) comes into play. Essentially it states that "best matches" is the model that has the highest likelihood, where likelihood for a dataset (measurements x_i) is defined as a function of the fit parameters as follows:

$$\mathcal{L}(\theta) = \prod_i PDF(x_i, \theta) \; . \tag{22}$$

In all its simplicity, it says that given data and a PDF one should choose the parameters (θ) of the PDF such that the likelihood ($\mathcal{L}$) is maximal. What is astounding is the variety of results that can be deduced from this principle (e.g. estimators in previous chapter).

Before applying this principle in more detail, let us just have a look at why it works at all. Imagine for example, that we have a series of repeated measurements x_i and are fitting these with a Gaussian distribution. If the mean was (far) off, then the value of the PDF would be rather small at each of the measurement values x_i and so would the resulting product and hence value of the likelihood $\mathcal{L}$. Thus, the mean should match the mean of the data to yield a higher likelihood value.

Likewise, if the Gaussian PDF is too narrow, it will of course yield high likelihood values for the central points, but those on the tail will have very small values, leading to a lower overall likelihood value $\mathcal{L}$. Conversely, if the Gaussian is too wide, some of the PDF will have significant values where there are no measurement points, again leading to a lower value of $\mathcal{L}$. Similarly for the normalisation.

Thus, the result is that the principle of maximum likelihood "forces" the PDF parameters to take on the values that best match the data. As it turns out, this definition of "best" is at the core of statistics, and the principle of maximum likelihood rules supremely. All the formulae for means, standard deviation, ChiSquare test (below), etc. can be derived from this principle.

In practice, it turns out that multiplying many numbers tends to lead to numerical problems, because the product either becomes very large or very small. For this reason, it is customary to take the logarithm of the likelihood (making the product a sum), and also multiplying by -2 to make it a minimisation problem. We'll get back to this factor two in a moment.

4.2 The ChiSquare method

The PDF of a measurement with uncertainties is always assumed Gaussian (if nothing else is stated). Thus, if we are fitting a series of measurements, that is x_i, y_i, and $\sigma(y_i)$, with a function $f(x, \theta)$, then ($-2\ln$ of) the likelihood would be:

$$\begin{aligned} -2\ln(\mathcal{L}(x,\theta)) &= -2\sum_i \ln\left(\frac{1}{\sqrt{2\pi}\sigma(y_i)} \exp\left(-\frac{1}{2}\left(\frac{y_i - f(x_i,\theta)}{\sigma(y_i)}\right)^2\right)\right) \\ &= -2\sum_i \ln\left(\frac{1}{\sqrt{2\pi}\sigma(y_i)}\right) - 2\sum_i -\frac{1}{2}\left(\frac{y_i - f(x_i,\theta)}{\sigma(y_i)}\right)^2 \\ &= C + \sum_i \left(\frac{y_i - f(x_i,\theta)}{\sigma(y_i)}\right)^2 . \end{aligned} \tag{23}$$

Now we want to minimise this with respect to the fit parameters θ (e.g. a and b, if fitting a line), which is why the first term can be considered a constant. It does not depend on the fit parameters θ, only the data points.

At this point, I hope the reader recognise the formula for the ChiSquare, which also explains the factor 2 in front of the negative log likelihood. When uncertainties are Gaussian (as they often are), minimising the likelihood is equivalent to minimising the ChiSquare. However, the ChiSquare comes with an advantage, namely a goodness-of-fit measure. That is, from the ChiSquare value (at the minimum) one can evaluate if the fit matches the data well. Sure, from minimising θ these are the parameters for which the fit function best fits the data, but that in itself does not tell you, if "best" is also good. It can be horrible. For the same reason, always inspect a fit visually, if you can, as your eyes are sharp at fitting and evaluating a fit ("Chi-by-eye" is the semi-technical term).

4.2.1 ChiSquare goodness-of-fit measure

So how to evaluate the ChiSquare value from a ChiSquare fit? Well, if —NOTE "IF"—your data is good and has correct Gaussian uncertainties, and if the model you fit with actually correctly describes the data (the ChiSquare assumptions), then each data point should contribute with a unit Gaussian number squared. And the distribution of such a sum is known: It is of course the ChiSquare distribution. The number of degrees of freedom ν for the ChiSquare distribution is the number of data points fitted, but adjusted for the number of fitting parameters:[3]

$$N_{\text{DoF}} = N_{\text{Data Points}} - N_{\text{Fit Parameters}} \ . \tag{24}$$

Note that even if there is no fit (i.e. when the model is given/fixed for example when comparing two histograms), one can still calculate the ChiSquare value, simply setting $N_{\text{Fit Parameters}} = 0$.

With the knowledge of how the ChiSquare value *should* distribute itself for a given situation (i.e. N_{DoF}), one can calculate **the probability of obtaining a certain ChiSquare value or worse** (the p value) from the integral of the ChiSquare distribution from the ChiSquare value obtained and to infinity (i.e. the survival function). If the assumptions are fulfilled, then these p-values will distribute themselves as a uniform distribution. This means, that when things are as they should be, all p-values are equally likely. However, when these assumptions are *not* fulfilled, in particular when the data does not follow the fit model, this increases the ChiSquare value, and in turn lowers the p-values obtained. The more statistically powerful the data (typically by high statistics and sharp observables), the more it diminishes. At some point it gets so low, that one can simply not uphold the hypothesis that the model matches the data.

The p-value of a ChiSquare probability can roughly be interpreted as follows:

If $0.01 < \textbf{Prob}(\chi^2, N_{\textbf{DoF}}) < 0.99$, then the fit is typically good. Where exactly to draw the limits can vary, but remember that if you do say 20 (independent) fits, where everything is perfectly in place (i.e. the conditions are fulfilled), then there should still on average be one p-value below 5%.

If $\textbf{Prob}(\chi^2, N_{\textbf{DoF}}) < 0.01$, then you have either been very unlucky (1:100) or (more likely) something is wrong. Assuming your data and uncertainties are perfect, then the model is unlikely.

If $0.99 < \textbf{Prob}(\chi^2, N_{\textbf{DoF}})$, then your fit is *too* good. The typical causes are overestimated uncertainties on the data or correlations between the points.

The calculation of a p-value assumes that the data is trustworthy, the uncertainties correct and Gaussian, and that the model matches the data (the "if" above).

> "It is however, more than doubtful whether the conditions for the theoretical validity of the method are realised in statistical practice, and the student would do well to regard the method as recommended chiefly by its comparative simplicity and by the fact that it has stood the test of experience."
>
> [G.U. Yule and M.G. Kendall 1958]

Having stood more than an additional half century, this is a testament to the robustness of the ChiSquare.

[3] This is related to the fact that a polynomial with N parameters can be made to go through N points.

EXAMPLE:

Consider a linear fit (2 parameters) to 9 data points, minimized to yield a ChiSquare value of 9.1. The number of degrees of freedom is $9 - 2 = 7$ and the calculated p-value (the probability of obtaining this χ^2 value or greater given the conditions) is Prob$(\chi^2 = 9.1, N_{\text{DoF}} = 7) = 0.246$. The conclusion is in this case is that the linear model fits the data well.
However, if the ChiSquare value had been 19.1, the p-value obtained would have been Prob$(\chi^2 = 19.1, N_{\text{DoF}} = 7) = 0.0079$, indicating that the model is unlikely to be matching the data. If the ChiSquare value instead had been 1.1, the p-value obtained would have been Prob$(\chi^2 = 1.1, N_{\text{DoF}} = 7) = 0.993$, pointing to the uncertainties being too large or a correlation between the points "damping" the expected statistical fluctuations.
Note that the p-value has a direct correspondence with "number of σs", in that a p-value of 0.0027 corresponds to a (two-sided) "3σ" observation.

4.2.2 *Versions of the ChiSquare calculation*

Several versions of the ChiSquare exists, mainly differing by how the uncertainty is calculated: Based on the Expected (E) or the Observed (O) number of events:

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \qquad \text{vs.} \qquad \sum_i \frac{(O_i - E_i)^2}{O_i} \,. \tag{25}$$

In both cases, the uncertainty is naturally taken to be the square root of the number of entries. In the first (Pearson) case all bins typically have non-zero values, and the question is "how many bins to include?". In the second case (implemented in Minuit) the ChiSquare sum can only be taken over the non-zero entries. This means that the fit doesn't "feel" the bins with zero entries, even if these contain valuable information (e.g. tails of distributions).
However, since the ChiSquare should only be used, when there is significant statistics (so that the uncertainties are Gaussian), the difference between these two approaches is rarely significant.

4.3 Binned likelihood fit

In addition to the unbinned likelihood fit, a binned version exists, building on the likelihood principle. Given a histogram where the number of observed (expected) entries are O_i (E_i), the binned likelihood is:

$$\mathcal{L}(\theta) = \prod_i \text{Poisson}(O_i, E(\theta)_i) = \prod_i \frac{E(\theta)_i^{O_i} e^{-E(\theta)_i}}{O_i!} \tag{26}$$

where the second equation is obtained simply by inserting into the Poisson distribution Eq. (17), which is the expected distribution for each bin. The advantage of an unbinned likelihood fit over a ChiSquare fit is, that the unbinned likelihood fit also works for low statistics. The downside is, that there is no simple goodness-of-fit test.

4.4 Which fitting method to use?

The unbinned likelihood fit is alluring, as it is the “best possible” in terms of requiring the least assumptions (e.g. Gaussian errors for the ChiSquare fit) and producing the results with the smallest uncertainties. But in terms of convergence, robustness, and not the least goodness-of-fit measure, the ChiSquare fit fares best.

The choice should thus reflect the situation. *If statistics is high* (e.g. all bins above 10 entries), and the Gaussian approximation of the uncertainties is thus valid, then the ChiSquare is recommendable due to its direct goodness-of-fit measure. *If statistics is low*, then the unbinned likelihood fit is probably the best choice.

In particle physics, the binned likelihood fits have been preferred as part of more complex fits across many samples and dimensions (with e.g. RooFit or HistFactory within). Since some channels are likely to have little statistics in some bins, the likelihood is chosen.

4.5 A Goodness-of-fit measure for likelihood fits

While likelihood values may take on essentially any range (also negative), there is no inherent way of testing if a likelihood fit actually yields a model that fits data well. However, it can be done, using simulation.

Imagine having performed a likelihood fit, obtaining the likelihood value $\mathcal{L}_{\text{fit}}$ and fit parameters $\hat{\theta}$. If the fit model reflects the data, then producing a new similar dataset based on the fit parameters should yield a new (simulated) dataset, for which the likelihood value should be similar. Repeating the simulation, fitting, and recording of the resulting likelihood value thus produces the expected distribution of the likelihood values, against which the likelihood value obtained from the fit to the true data $\mathcal{L}_{\text{fit}}$ can be compared.

5 Hypothesis Testing

Suppose in a beer tasting, that someone gets 9 out of 10 right. Does that prove that the person can taste the differences between beers? The slightly surprising answer is “No”.

What we can say is that the result is inconsistent (at some significance level) with the hypothesis that the person chooses at random. This leaves us with the alternative hypotheses, that the person can taste the differences or has cheated (consciously or unconsciously).

In statistics one can never prove a hypothesis directly. However, one can set up alternative hypotheses and disprove these. That is how one works in statistics...

5.1 Nomenclature in Hypothesis Testing

Hypothesis testing is like a criminal trial. The basic “null” hypothesis is Innocent (denoted H_0) and this is the hypothesis we want to test, compared to an “alternative” hypothesis, Guilty (denoted H_1). Innocence (“negative”) is initially assumed, and this hypothesis is only rejected, if enough evidence proves otherwise, i.e. that the probability of innocence is very small (“beyond reasonable doubt”).

Given two possible truths (innocent or guilty) and two possible verdicts (acquittal or conviction), there are four outcomes: True positive (TP, guilty and convicted), false positive (FP, innocent but convicted), false negative (FN, guilty but acquitted), and true negative (TN, innocence and acquitted). Given these four numbers, one can determine two rates:

False Positive Rate (FPR) is defined as $FPR = FP/(FP + TN)$ is the probability of rejecting H_0, when it is true (e.g. the rate of convicting the innocents).

False Negative Rate (FNR) is defined as $FNR = FN/(FN+TP)$ is the probability of accepting H_0, when it is false (e.g. the rate of acquitting the guilty).

For a given test, these two rates are inter-dependent. Take for example electron identification: If you lower the threshold for being an electron H_0, then you will also lower the FPR (rejecting a true electron), but you will at the same time increase the FNR (accepting a non-electron). Similarly, TPR and TNR can be defined.

The purpose of a hypothesis test is to yield (calculable/predictable) distributions of a test statistic t for the Null (H_0) and Alternative (H_1) hypotheses, which are as separated from each other as possible (in order to minimise FPR and FNR). The way to determine the separation is to plot the TPR against the FPR as a function of the possible selection criteria. This produces the Receiver Operating Characteristic (ROC) curve, which incorporates essentially everything about a given test.

Take again the court example. If we required impossible amounts of evidence, then no one would be convicted, and both the FPR (the innocent) and the TPR (the guilty) would be 0. Conversely, if we required no evidence at all, then everyone would be convicted, and the FPR and TPR would be 1. However, somewhere in between these extremes lies the power of the courts: Obtaining a high TPR while maintaining a low FPR.

Note that there can be several hypothesis, and that hypothesis testing may exclude any number of hypothesis ranging from none to all! Testing multiple hypothesis simultaneously is more complex.

5.2 Steps in Hypothesis Testing

Consider a case for which you want to do a hypothesis testing, and state a null hypothesis along with an alternative hypothesis. Think about the statistical assumptions you are making (independence, distributions, etc.), and then decide for an appropriate statistical test. Then define the relevant test statistic t, which could be anything from a counting number to a machine learning output. Next, derive/calculate the test statistic distribution under null and alternative hypothesis. In standard cases, these are well known distributions (Poisson, Gaussian, Student's t, etc.).

Before you do the actual test, you should select a significance level (α) that is a probability threshold below which the null hypothesis will be rejected. In particle physics we typically use 0.0027 (3σ) for "evidence", 0.000063 (4σ) for "observation", and 0.00000057 (5σ) for "discovery". Other sciences tend to use slightly lower values of α (e.g. 0.05 (2σ) in biology and medicine). There is no universally accepted value, as there should not be as "the weight of evidence for an extraordinary claim must be proportioned to its strangeness" [Pierre-Simon Laplace]. For an excellent discussion of this related to particle physics see Louis Lyons, "Discovering the significance of 5σ".

Once all the pieces above are in place, compute from (otherwise blinded) observations/data the value of the test statistic t, and use this to calculate the probability of observation under null hypothesis (p-value). Reject the null hypothesis for the alternative if the p-value is below the significance level.

5.3 The Neyman–Pearson lemma and Wilk's theorem

While a single likelihood value says little, the ratio of likelihood values (after minimisation) between two competing hypothesis (the null $\mathcal{L}_0$ and the alternative $\mathcal{L}_1$) plays a central role in hypothesis testing. If the two hypothesis are simple (i.e. have no free parameters) then the Neyman–Pearson Lemma (loosely) states that the $(-2\ln)$ likelihood ratio D defined as:

$$D = -2\ln\left(\frac{\mathcal{L}_0}{\mathcal{L}_1}\right) = -2\ln\mathcal{L}_0 + 2\ln\mathcal{L}_1 \tag{27}$$

is the best possible test statistic that exists. Now, that is a strong statement, which is why likelihood ratios are often used. An example use case is the determination of the Higgs particle spin, where the null hypothesis was 0, while the alternative hypothesis was 2 (for theoretical reasons spin 1 is not (well, hardly) possible). In this case none of the hypothesis have any free parameters, and the lemma applies. The challenge lies in determining the expected distributions of D for the two hypothesis.

Most often though, the hypotheses and associated likelihoods have free parameters. If the null hypothesis is nested in the alternative hypothesis (i.e. that the alternative hypothesis contains the null hypothesis), then the very nice Wilk's Theorem states that in the limit of large statistics the likelihood ratio D approximately follows a Chi-Square distribution with $N_{\text{DoF}} = N_{\text{DoF}}() - N_{\text{DoF}}()$, if H_0 is true.

The reason for the nested requirement is that it ensures that the alternative hypothesis will always have a higher likelihood value than the null, and hence D will always be positive, as it should be to reflect a ChiSquare value. If uncertainties are Gaussian (e.g. high statistics histograms) then Wilk's theorem extends to include Chi-Square differences also.

EXAMPLE:

Imagine that you are fitting a (high statistics) data sample which is an exponentially falling spectrum of background events with a potential Gaussian peak in the middle. In order to test the null hypothesis H_0 that there is only background (2 parameter fit) against the alternative hypothesis H_1 that is a Guassian peak in addition (5 parameter fit), you fit the data for each case.

The resulting fits (without and with a Gaussian peak component) yields likelihood values of $-2\ln\mathcal{L}_0 = -10017$ and $-2\ln\mathcal{L}_1 = -10000$. The likelihood values in themselves do not reveal much, but their difference yields $D = 17$ for $5 - 2 = 3$ degrees of freedom, which results in a p-value of 0.00071 (approximately 3.2σ for a one-sided test).

You thus conclude, that you have found evidence for a peak.

5.4 Various hypothesis tests

Many different types of statistical tests exists. Below a few of the more frequently used ones are listed.

5.4.1 One-sample test

Used when comparing e.g. the mean of a sample to a known value. Example: Comparing mean of measurements to known constant: $\mu_{\text{exp}} = 2.91 \pm 0.02$ vs. $c = 2.99$, which is thus a 4σ difference.

5.4.2 *Two-sample test*

Used when comparing e.g. the means of two samples. If the samples are high statistics (or the sample σs are known), then the test works much like the one-sample test. Example: Comparing a sample to control sample: $\mu_{\text{exp}} = 4.01 \pm 0.12$ vs. $\mu_{\text{control}} = 3.88 \pm 0.05$, which is thus a $(4.01 - 3.88)/\sqrt{0.12^2 + 0.05^2} = 1.0\sigma$ difference.

In case there is a pairing between the two samples (e.g. twins), this may reduce the test to a paired test, which is a one-sample test, the advantage being that the pairing will cancel out much of the variance from nuisance parameters (e.g. genetic biases in the case of twins).

5.4.3 *Chi-squared test*

This test evaluates the adequacy of a model compared to data (or between two datasets) through the χ^2 value. Example: Model fitted to (possibly binned) data, yielding p-value = Prob($\chi^2 = 45.9, N_{\text{DoF}} = 36) = 0.125$.

More generally, this test can be used for determining parameters of and over-determined system of equations (i.e. with more equations than unknowns), which is a powerful way of solving many averaging, calibration, and combinatorial challenges.

5.4.4 *Wald–Wolfowitz runs test*

The WW runs test is a binary check for independence between entries in a series of values. Imagine a fit, where you want to check if the N data points lie randomly above and below the fit or rather in collected groups (islands). This can be tested, since the number of expected islands $\mu(N_{\text{runs}})$ and its variation $\sigma(N_{\text{runs}})$ can be determined from the number of entries passing some binary requirement (e.g. above or below fit) N_+ and N_-:

$$\mu(N_{\text{runs}}) = 1 + \frac{2N_+N_-}{N}\ , \tag{28}$$

$$\sigma(N_{\text{runs}}) = \sqrt{\frac{2N_+N_-(2N_+N_- - N)}{N^2(N-1)}}\ . \tag{29}$$

In a sense, the WW runs test is complimentary to the ChiSquare test in that it considers the sign rather than the size of a deviations. As such, it serves well to check a fit, where the result despite a good ChiSquare p-value (possibly due to overestimated systematic uncertainties) does not seem to follow the trends of the data well. Note that $\sigma(N_{\text{runs}})$ is only approximately Gaussian, when N_+ and N_- are greater than about 10.

Example: You fit $N = 25$ data points, and the resulting distribution of data points above/below the fit is $+++++-------++++++------$++. Thus $N_+ =$ 13 and $N_- =$ 12, which yields $\mu =$ 13.5 $\pm$ 2.4 islands. You observe $\mu(N_{\text{runs}}) =$ 5, which is suspiciously ((13.5 $-$ 5)/2.4 $= 3.5\sigma$) low.

5.4.5 *Kolmogorov–Smirnov test*

This test compares the degree to which two 1D distributions are compatible, i.e. a p-value for the distributions being samples of the same underlying PDF. Example: Compatibility between data and MC sample, yielding p-value = 0.87 (thus the distributions are compatible). The Kolmogorov–Smirnov test does not make any assumption about distributions, which makes it very powerful.

6 Short note on setting limits

If you observe 0 events of type X in an experiment, then where to set the limit? The way you should think about it is as follows: The number of X-events is Poisson distributed, because you probably made many attempts (N) to produce and observe X-events, but the probability of doing so (p) was rather low. If this is the case, then how large could the value of $\lambda = Np$ be, such that it is still consistent to have say 5% chance of observing N = 0?

The answer is obtained from increasing λ until the observed is Poisson$(N_{\text{obs}} = 0, \lambda = N_{\text{obs}} = 2.996) = 0.05$. Thus the 95% Upper Limit is (rounded to be) 3.0. In case you wanted the 90% Upper Limit, the answer would be 2.30.

7 Final words of advice

Statistics and the general understanding of numbers and their relations play a central role in particle physics. Embrace it! Doing so makes you a member of the much sought-after class of persons, who are hyper literate in numbers. This skill will take you far, not just in particle physics but in most endeavours encompassing numericals.

At times you will of course encounter difficult situations, where you are unsure of what the correct statistical approach is. You can of course test this numerically with simulations for closure tests, but the doubt might remain. This happens to all experts in all fields, the author included. In that case, do not be too troubled, but rather lean on the wise words of a mentor and fellow statistician:

> “Don’t worry too much about statistics. Just tell us what you do and do what you tell us.”
>
> [Roger Barlow, ICHEP conference 2006 in Moscow]

Neutrino physics

Gabriela Barenboim[a]
[a]University of Valencia and IFIC, Valencia, Spain

Surpassing all expectations, the Standard Model has predicted the outcomes of nearly every experiment conducted thus far. Neutrinos have no mass in it. However, we have gathered strong evidence in the last twenty years suggesting that the neutrinos have small but non-zero masses, These masses are a real delight because they allow neutrinos to oscillate and change flavor. I go over the characteristics of neutrinos both inside and outside the Standard Model in these lectures, as well as their incredible potential. I also revise the bits of data that defy the standard picture of the three neutrinos and discuss the possibilities of employing neutrinos to uncover any physics hidden outside the Standard Model.

1 Introduction

In the last twenty years, neutrino physics underwent a drastic change. Neutrinos have non-zero masses, which implies that leptons mix—this is an unequivocal finding. Neutrinos can change from one state, or flavor, to another, as experimental evidence has shown. All knowledge that we have about neutrinos is relatively recent, younger than thirty years of age. Since neutrino physics is still in its early stages as a solid science, it is in a wild and extremely exciting (and excited) state, similar to any young adult. But let's first discuss how and why neutrinos were born, before diving into the late "news" about them.

Several holy cows met their demise in the 1920s, and physics was no different. It appeared that the subatomic realm defies one of physics' most cherished principles: energy conservation. A non-negligible fraction of the energy of some radioactive nuclei seemed to just disappear, leaving no sign of its existence.

This chapter should be cited as: Neutrino physics, Gabriela Barenboim, DOI: 10.23730/CYRSP-2025-009.123,
in: Proceedings of the 2023 European School of High-Energy Physics,
CERN Yellow Reports: School Proceedings, CERN-2025-009, DOI: 10.23730/CYRSP-2025-009, p.123.

"Dear radioactive Ladies and Gentlemen, . . . ," Pauli wrote in a quasi-apologetically worded letter to a meeting in 1920 (by now famous [1])."... as a desperate remedy to save the principle of energy conservation in beta decay, ... I propose the idea of a neutral particle of spin half". Pauli proposed that an additional particle—one that lacked an electric charge, mass, and was impossible to detect, and therefore invisible, because it was only very weakly interacting—was responsible for absorbing the missing energy.

It wasn't long before Fermi proposed the four-Fermi Hamiltonian to explain beta decay using the neutrino, electron, neutron, and proton. With these characteristics, the neutrino was thus introduced as one of the few components of the particle zoo. Weak interactions were then born, took center stage and never left, giving rise to a new field. To close the loop, Cowan and Reines obtained the experimental signature of anti-neutrinos emitted by a nuclear power plant twenty years after Pauli's letter.

Weak interactions gained credibility as a real new force of nature, with the neutrino being a fundamental component, as more particles involved in them were discovered in the years after the discovery of the neutrino.

Over the ensuing years, additional experimental testing revealed that there were, in fact, three different types, or "flavors," of neutrinos (named for the charged lepton they were produced in conjunction with: electron neutrinos (ν_e), muon neutrinos (ν_μ), and tau neutrinos (ν_τ)), and that, to the extent that we could test, had no mass (and no charge) at all.

A new test using neutrinos from the sun revealed that the neutrino saga was just getting started, even though it could have easily ended there.

In the original Standard Model, neutrinos were completely massless and as a consequence were naturally flavor eigenstates,

$$\begin{aligned} W^+ \longrightarrow e^+ \,+\, \nu_e \quad &; \quad Z \longrightarrow \nu_e \,+\, \bar{\nu}_e \;; \\ W^+ \longrightarrow \mu^+ \,+\, \nu_\mu \quad &; \quad Z \longrightarrow \nu_\mu \,+\, \bar{\nu}_\mu \;; \\ W^+ \longrightarrow \tau^+ \,+\, \nu_\tau \quad &; \quad Z \longrightarrow \nu_\tau \,+\, \bar{\nu}_\tau \;. \end{aligned} \tag{1}$$

In fact they would have been flavor eigenstates even if they were massive, if they had shared the same mass. In any case, they were supposed to move at the speed of light precisely because they were massless. However, their masslessness not only established their propagation speed but also fixed their flavor as they moved. It follows that, in terms of flavor, zero mass neutrinos were not a compelling subject for research, particularly when compared to quarks.

However, if neutrinos were massive, and these masses were not degenerate, as we have mentioned degenerate masses flavor-wise are identical to the zero mass case, would mean that neutrino mass eigenstates exist $\nu_i, i = 1, 2, \ldots$, each with a mass m_i. The effect of leptonic mixing is transparent when considering the leptonic decays of the charged vector boson W , $W^+ \longrightarrow \nu_i + \overline{\ell_\alpha}$. Where, $\alpha = e, \mu$, or τ, and ℓ_e refers to the electron, ℓ_μ the muon, or ℓ_τ the tau.

We denominate ℓ_α as the charged lepton of flavor α. Mixing basically implies that when the charged boson W^+ decays to a given kind of charged lepton $\overline{\ell_\alpha}$, the neutrino that goes along is not generally the same mass eigenstate ν_i. *Any* of the different ν_i can appear. Or all of them!!

The amplitude for the decay of a vector boson W^+ to a particular mix $\overline{\ell_\alpha} + \nu_i$ is given by $U^*_{\alpha i}$.

The neutrino that is emitted in this decay alongside the given charged lepton $\overline{\ell_\alpha}$ is then

$$|\nu_\alpha >= \sum_i U^*_{\alpha i} \, |\nu_i > \quad . \tag{2}$$

This specific mixture of mass eigenstates yields the neutrino of flavor α.

As there are nine of these elements, not independent, the different $U_{\alpha i}$ can be collected in a unitary matrix, in the same way they were collected in the CKM matrix in the quark sector. This matrix receives the name of the leptonic mixing matrix, or U_{PNMS} [2]. The unitarity of U ensures that every time a neutrino of flavor α produces a charged lepton through its interaction, the produced charged lepton will always be ℓ_α, the charged lepton of flavor α. That is, a ν_e produces exclusively an e, a ν_μ exclusively a μ, and similarly ν_τ can only make a τ.

Any mass eigenstate ν_i can be represented as an analogous linear combination of the three flavors by simply inverting the expression (2), which represents each neutrino of a given flavor as a linear combination of the three mass eigenstates:

$$|\nu_i >= \sum_\alpha U_{\alpha i} \, |\nu_\alpha > \quad . \tag{3}$$

The α-fraction or $|U_{\alpha i}|^2$ is clearly the amount of α-flavor in ν_i. This α-content, also known as the fraction, expresses the likelihood that a charged lepton produced by the interaction with a ν_i will have α flavor.

2 Neutrino oscillations basics

Let's begin with an explanation of what we understand for the phenomena known as neutrino flavor transitions, or oscillation for short: Together with a (positively) charged lepton $\overline{\ell_\alpha}$ of flavor α, a source produces or emits a neutrino. In this sense, the neutrino does have a distinct flavor at the emission site—a ν_α, that is. The neutrino then propagates, i.e. travels a distance L, until it is absorbed.

By now, the neutrino has (sometimes) reached the detector. These interactions produce another charged lepton, ℓ_β, of flavor β, which we can detect. This allows us to determine that the neutrino at the target is, once more a ν_β, a neutrino with a distinct flavor. Naturally, not every time both flavors are the same. Only sometimes, $\beta \neq \alpha$ For example, if ℓ_α is μ but ℓ_β is τ. Then, when traveling from the source to the detection point, the neutrino changes from ν_α to ν_β.

The transition from one flavor to another, $\nu_\alpha \longrightarrow \nu_\beta$, is not unique to neutrinos, but is just one example of the quantum mechanical effect known to exist in two level systems.

From Eq. (2), ν_α is the coherent superposition of the three mass eigenstates, ν_i, the actual neutrino traveling from the moment of its creation to its detection. So it can be any one of the three ν_is. That is why the contribution of each ν_i should be included coherently in the equation. As a result, the transfer energy Amp($\nu_\alpha \longrightarrow \nu_\beta$) receives a contribution from ν_i and appears as a product of three phases. The first component is the probability amplitude of the neutrino produced at the formation point with the $\overline{\ell_\alpha}$ lepton to be any of the three mass eigenstates. In particular: say a ν_i and is given by $U^*_{\alpha i}$.

The second part of our results is the amplitude of ν_i generated by the source to cover the distance to the detector. Now, let's name this element Prop(ν_i) and postpone the calculation of its value until later. The last (third) component is the probability amplitude of the traveling neutrino ν_i to produce a charged lepton ℓ_β .

As probabilities must be conserved, we know that the Hamiltonian that describes the interactions between neutrinos, charged leptons and charged bosons W bosons has to be hermitian, therefore Amp($W \longrightarrow \overline{\ell_\alpha} \nu_i) = U^*_{\alpha i}$, then Amp $(\nu_i \longrightarrow \ell_\beta W) = U_{\beta i}$. Thus, the last piece of the product, the ν_i contribution, is given by $U_{\beta i}$, and

$$\text{Amp}(\nu_\alpha \longrightarrow \nu_\beta) = \sum_i U^*_{\alpha i} \ \text{Prop}(\nu_i) \ U_{\beta i} \ . \tag{4}$$

The value of Prop(ν_i) is still up for determination. We'd best examine the ν_i in its rest frame in order to ascertain it. In such a framework, we shall designate the time as τ_i. In this frame of reference, ν_i's state vector fulfills the Schrödinger equation that, if it does have a rest mass m_i, can be written as

$$i \frac{\partial}{\partial \tau_i} |\nu_i(\tau_i) > = m_i |\nu_i(\tau_i) > \quad . \tag{5}$$

whose solution is given by

$$|\nu_i(\tau_i) > = e^{-i m_i \tau_i} |\nu_i(0) > \quad . \tag{6}$$

After some time the initial ν_i, $|\nu_i(0) >$, has become the evolved state $|\nu_i(\tau_i) >$, specifically $\exp[-i m_i \tau_i]$. Thus, the amplitude for a particular mass eigenstate ν_i to move freely throughout a time τ_i is just the amplitude $< \nu_i(0)|\nu_i(\tau_i) >$. Therefore, Prop($\nu_i$) only includes this amplitude in which we have utilized the fact that τ_i, the proper time, is the amount of time required for ν_i to travel the distance between the source and the detector.

However, if we want Prop(ν_i) to be useful to us, we must first re-write it in terms of variables that we can measure, which means expressing it in variables in the laboratory frame. A natural choice is the distance L that the neutrino travels between the source and the detector, shown in the laboratory frame and the time t, elapsed during the journey in the lab frame. The distance L is determined by the experiment by selecting the source and detector resolution points, which are unique for each test setup. Similarly, the value t is determined by the experiment by choosing when the neutrinos appear and when the neutrinos die, or are detected. Thus, L and t are determined by the test settings, which are the same for all ν_i in the beam. Different ν_i cover the same distance L in the same time t.

Two more lab frame variables, the energy E_i and three momentum p_i of the neutrino mass eigenstate ν_i, need to be found. The expression for the $m_i \tau_i$ appearing in the ν_i propagator Prop(ν_i) may be obtained in terms of the (simple to measure) lab frame variables by leveraging the Lorentz invariance of the four component internal product (scalar product),

$$m_i \tau_i = E_i t - p_i L \ . \tag{7}$$

At this point, one could counter that the time t that elapses from the moment a neutrino is created until it dies in the detector is not truly observed because neutrino sources in real life are essentially constant in time. This is a perfectly valid argument. Actually, an experiment spreads out over the time t that the neutrino needs to travel through. But let's assume that the neutrino signal generated in the detector is composed coherently of two components of the neutrino beam: the first, with energy E_1, and the second, with energy E_2 (both measured in the lab frame). Let's now refer to the time t as the neutrino's travel over the distance between production and detection points.

The component with energy E_j $(j = 1, 2)$ has then picked up a phase factor $\exp[-iE_j t]$ by the time it reaches the detector. Consequently, there will be an interference with a phase factor of $\exp[-i(E_1 - E_2)t]$ between the constituents of the E_1 and E_2 beams. This factor vanishes when smeared throughout the non-observed journey time t, *except when* $E_2 = E_1$. Consequently, the neutrino oscillation signal is only coherently contributed to by components of the neutrino beam that have the same energy [3]. Particularly, only the beam's various mass eigenstate components with equivalent energy weigh in. Everything else is averaged out.

A mass eigenstate ν_i with mass m_i and energy E has three momentum p_i, whose absolute value is given, thanks to its dispersion relation, by

$$p_i = \sqrt{E^2 - m_i^2} \cong E - \frac{m_i^2}{2E} \quad . \tag{8}$$

Since the neutrinos' masses are pitifully small and given the typical energies E that are involved in any experiment, we have used the fact that $m_i^2 \ll E^2$, i.e. the lowest energy neutrinos have MeV energies while its masses are sub-eV at most. It is simple to show that for a given energy E, the phase $m_i \tau_i$, appearing in Prop(ν_i), takes the value indicated by Eqs. (7) and (8),

$$m_i \tau_i \cong E(t - L) + \frac{m_i^2}{2E} L \quad . \tag{9}$$

When computing the transition amplitude, the phase $E(t - L)$ will finally vanish because it appears in all the interfering terms. The common phase factor, after all, has an absolute value of one. As a result, we can discard it right now and use

$$\text{Prop}(\nu_i) = \exp[-im_i^2 \frac{L}{2E}] \quad . \tag{10}$$

Entering this into Eq. (4), we easily find that the amplitude for a neutrino that starts out as a ν_α and travels L with energy E to be detected as a ν_β is given by

$$\text{Amp}(\nu_\alpha \longrightarrow \nu_\beta) = \sum_i U_{\alpha i}^* \, e^{-im_i^2 \frac{L}{2E}} U_{\beta i} \quad . \tag{11}$$

As long as the neutrinos pass through vacuum, the statement above holds true for any number of neutrino flavors and an equal number of mass eigenstates. Squaring it yields the probability P($\nu_\alpha \longrightarrow \nu_\beta$) for $\nu_\alpha \longrightarrow \nu_\beta$.

$$\begin{aligned} &\mathrm{P}(\nu_\alpha \longrightarrow \nu_\beta) = |\text{Amp}(\nu_\alpha \longrightarrow \nu_\beta)|^2 = \delta_{\alpha\beta} - \\ &4 \sum_{i>j} \Re(U_{\alpha i}^* U_{\beta i} U_{\alpha j} U_{\beta j}^*) \sin^2\left(\Delta m_{ij}^2 \frac{L}{4E}\right) + 2 \sum_{i>j} \Im(U_{\alpha i}^* U_{\beta i} U_{\alpha j} U_{\beta j}^*) \sin\left(\Delta m_{ij}^2 \frac{L}{2E}\right), \end{aligned} \tag{12}$$

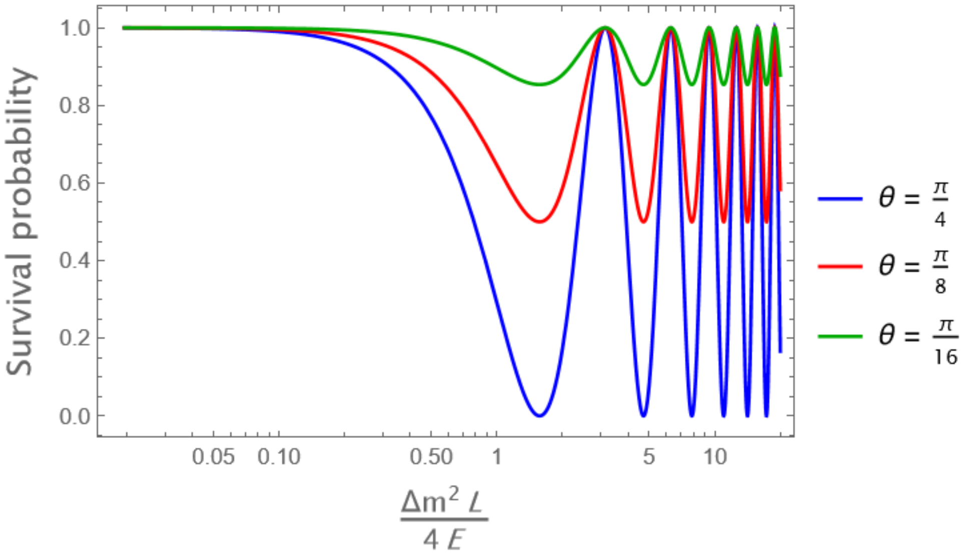


Fig. 1: Survival probability as a function of the kinematic phase. The amplitude of the oscillation is given by the mixing angle. The kinematic phase has to be order one for the oscillations to be seen.

with $\Delta m^2_{ij} \equiv m^2_i - m^2_j$. We have utilized the fact that the mixing matrix U is unitary to obtain Eq. (12).

Since the oscillating neutrino was created along with a charged *antilepton* $\bar{\ell}$ and gives birth to a charged *lepton* ℓ once it reaches the detector, the oscillation probability $\mathrm{P}(\nu_\alpha \longrightarrow \nu_\beta)$ that we have just obtained corresponds to that of a *neutrino* rather than a *antineutrino*. By utilizing the fact that the two transitions, $\overline{\nu_\alpha} \longrightarrow \overline{\nu_\beta}$ and $\nu_\beta \longrightarrow \nu_\alpha$, are CPT conjugated processes, the corresponding probability $\mathrm{P}(\overline{\nu_\alpha} \longrightarrow \overline{\nu_\beta})$ for an antineutrino oscillation can be found from $\mathrm{P}(\nu_\alpha \longrightarrow \nu_\beta)$. Thus, assuming that neutrino interactions respect CPT [4],

$$\mathrm{P}(\overline{\nu_\alpha} \longrightarrow \overline{\nu_\beta}) = \mathrm{P}(\nu_\beta \longrightarrow \nu_\alpha) \quad . \tag{13}$$

Then, it is evident that $\mathrm{P}(\overline{\nu_\alpha} \longrightarrow \overline{\nu_\beta})$ and $\mathrm{P}(\nu_\alpha \longrightarrow \nu_\beta)$ will not be equal in general if the mixing matrix U is complex. $\mathrm{P}(\overline{\nu_\alpha} \longrightarrow \overline{\nu_\beta}) \neq \mathrm{P}(\nu_\alpha \longrightarrow \nu_\beta)$ would provide evidence of CP violation in neutrino oscillations, since $\overline{\nu_\alpha} \longrightarrow \overline{\nu_\beta}$ and $\nu_\alpha \longrightarrow \nu_\beta$ are CP conjugated processes. Since CP violation has only been detected in the quark sector up to this point, measuring it in neutrino physics would be quite interesting.

We have been operating in natural units thus far. An observation made clear by examining the dispersion relation Eq. (9), where we have happily set both the c and $\hbar$ factors to one. If we return them to the oscillation probability, we discover that

$$\sin^2\left(\Delta m^2_{ij} \frac{L}{4E}\right) \quad \longrightarrow \quad \sin^2\left(\Delta m^2_{ij} c^4 \frac{L}{4\hbar c E}\right) \tag{14}$$

After that, investigating the semi-classical limit, $\hbar \longrightarrow 0$, is simple and informative. The oscillation averages to 1/2 in this limit, and the oscillation length and phase both go to 0 and infinity, respectively. There is no longer any interference pattern. When we let the mass difference Δm^2 grow, we have a similar scenario. This is precisely the quark sector's behavior, and the reason why, despite our knowledge that mass eigenstates and flavor eigenstates do not coincide, we never investigate quark oscillations.

Given actual units, which are not "natural" units, the oscillation phase can be expressed as follows:

$$\Delta m_{ij}^2 \frac{L}{4E} = 1.27\, \Delta m_{ij}^2 (\mathrm{eV}^2) \frac{L\,(\mathrm{km})}{E\,(\mathrm{GeV})} \quad . \tag{15}$$

consequently, given that $\sin^2[1.27\, \Delta m_{ij}^2(\mathrm{eV}^2) L\,(\mathrm{km})/E\,(\mathrm{GeV})]$ can only be experimentally detected, i.e. not smeared out, if its argument is roughly around one. Thus, an experimental setup that has a baseline of L (km) and an energy of E (GeV), is sensitive to the neutrino mass squared difference $\Delta m_{ij}^2(\mathrm{eV}^2)$ of order $\sim [L\,(\mathrm{km})/E\,(\mathrm{GeV}]^{-1}$.

To investigate mass differences Δm_{ij}^2 down to $\sim 10^{-4}$ eV^2, for instance, an experiment with a baseline of $L \sim 10^4$ km, or nearly the size of Earth's diameter, and $E \sim 1$ GeV should be conducted. This fact demonstrates that even pitifully small neutrino mass variations can be tested by neutrino long-baseline experiments. It achieves this by taking advantage of the quantum mechanical interference between amplitudes whose relative phases are provided exactly by these minuscule neutrino mass discrepancies. By selecting L/E properly, these amplitudes can be converted into substantial impacts.

But let's analyze the oscillation probability further and see if we can learn more about neutrino oscillations by studying its mathematical formula. It's clear from $\mathrm{P}(\overset{(-)}{\nu}_\alpha \longrightarrow \overset{(-)}{\nu}_\beta)$ that if the mass of all the neutrinos is zero, or they are all mass degenerate, that is, if all $\Delta m_{ij}^2 = 0$, then, $\mathrm{P}(\overset{(-)}{\nu}_\alpha \longrightarrow \overset{(-)}{\nu}_\beta) = \delta_{\alpha\beta}$.

Neutrinos are therefore massive, and their masses are not degenerate, as evidenced by the experimental finding that they can change from one flavor to another. In actuality, it was this very evidence that established the mass of neutrinos beyond a shadow of a doubt.

All observed oscillations of neutrinos have at some point involved neutrinos passing through matter. However, the expression we deduced is limited to flavor change in vacuum and ignores any interaction that may occur between the neutrinos and the matter they pass through en route to their detector. Therefore, the question still stands whether the observed flavor transitions could actually be caused by some unidentified flavor-changing interaction between neutrinos and matter rather than neutrino masses.

In response to this inquiry, a few points ought to be made. Foremost, while it is true that the Standard Model of elementary particle physics only includes massless neutrinos, it also describes all the possible ways a neutrino can interact, and does so in an extraordinarily well-corroborated way. A description that however does not include flavor change.

Second, matter effects are predicted to be pitifully small for some of the experimentally observed processes where neutrinos do change flavor. In those cases, however, the evidence suggests that the flavor transition probability depends on L and E through the combination L/E, as predicted by the oscillation hypothesis. Besides, L/E is the exact proper time that passes in the neutrino's rest frame while it travels a distance L with energy E, modulo a constant. Hence, rather than being the outcome of a reaction with matter, these flavor transitions behave as though they were a real evolution of the neutrino over time.

Let's now investigate the scenario in which the leptonic mixing is negligible. This would suggest that the emerging charged antilepton $\overline{\ell_\alpha}$ of flavor α always follows the *same* neutrino mass eigenstate ν_i in the charged boson decay $W^+ \longrightarrow \overline{\ell_\alpha} + \nu_i$, which as we established has an amplitude $U_{\alpha i}^*$. In other words, if $U^*_{\alpha i} \neq 0$, then for all $j \neq i$, $U_{\alpha j}$ becomes zero because of unitarity. Consequently, it is evident from Eq. (12) that $\mathrm{P}(\overset{(-)}{\nu_\alpha} \longrightarrow \overset{(-)}{\nu_\beta}) = \delta_{\alpha\beta}$. As neutrinos are known to change flavor, this undoubtedly indicates the presence of a non-trivial mixing matrix.

Consequently, there are essentially two methods left for detecting neutrino flavor change. The first is to notice that some neutrinos of a new flavor β which differs from the original flavor α are present in a beam of neutrinos that were all formed with the same flavor, let's say α. We refer to this as appearance experiments. The second method involves beginning with a beam of identical ν_αs, whose flux is either known or measured, and seeing that this flux is exhausted after a certain distance. These kinds of investigations are known as disappearance studies.

The transition probability in vacuum depends on L/E and oscillates with it, as demonstrated by Eq. (12). This is the reason behind the term "neutrino oscillations" for neutrino flavor transitions. Note that the squared-mass *differences* determines the neutrino transition probabilities rather than the individual neutrino masses or masses squared. Thus, the neutrino mass squared spectrum is the only quantity that oscillation experiments can measure. not on an absolute scale. The pattern can be tested by experiments, but the distance above zero at which the entire spectrum lies cannot be found.

It is evident that neutrino transitions only change the distribution of flux among the many flavors in a neutrino beam, not its total flux. In fact, given the unitarity of the U matrix and Eq. (12), it is clear that

$$\sum_{\beta} \mathrm{P}(\overset{(-)}{\nu_\alpha} \longrightarrow \overset{(-)}{\nu_\beta}) = 1 \ \ , \tag{16}$$

where the total includes the original flavor α as well as all other flavors β. It is evident from Eq. (16) that the likelihood of a neutrino changing its flavor, when paired with the probability of it being unchanged at birth, equals one. Thus, flavor changes don't affect the overall flow. However, some of the flavors that a neutrino can oscillate into, $\beta \neq \alpha$, might be *sterile* flavors—that is, flavors that avoid detection by not participating in weak interactions. An experiment that measures the total *active* neutrino flux—that is, the flux related to those neutrinos that couple to the weak gauge bosons: ν_e, ν_μ, and ν_τ—will note that it is smaller than the original flux if any of the original (active) neutrino flux becomes sterile. No flux has ever been overlooked in any experiment up until this point.

The various mass eigenstates ν_i that coherently contribute to a beam are typically assumed to share the same *momentum* in literature descriptions of neutrino oscillations, rather than the same *energy,* as we have proved they must have. Although the assumption of equal momentum is incorrect in theory, it is irrelevant (or not a mistake worth worrying about) because it leads to the same oscillation probability as the ones we have found, as can be readily demonstrated.

The case where just two flavors engage in the oscillation is a pertinent and intriguing example of the (not so simple) formula for $\mathrm{P}(\overline{\nu_\alpha} \longrightarrow \overline{\nu_\beta})$. Numerous experiments are quite rigorously described by the only-two-neutrino scenario. Actually, a more complex (three neutrino description) was only required recently (and in a few experiments) in order to fit observations.

In order to ensure that only one squared-mass difference, $m_2^2 - m_1^2 \equiv \Delta m^2$, surfaces, let's assume that only two mass eigenstates, which we will name ν_1 and ν_2, and two reciprocal flavor states, which we will name ν_μ and ν_τ, are important. Furthermore, the mixing matrix U can be expressed as follows by ignoring phase variables that are demonstrably insignificant to oscillation probabilities:

$$\begin{pmatrix} \nu_\mu \\ \nu_\tau \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \tag{17}$$

As a 2×2 rotation matrix, the unitary mixing matrix U of Eq. (17) is parameterized by a single rotation angle θ, which is known as the mixing angle in neutrino physics. We can easily demonstrate that, for $\beta \neq \alpha$, when only two neutrinos are important, by plugging the U of Eq. (17) and the unique Δm^2 into the general formula of the transition probability $\mathrm{P}(\overline{\nu_\alpha} \longrightarrow \overline{\nu_\beta})$

$$\mathrm{P}(\overline{\nu_\alpha} \longrightarrow \overline{\nu_\beta}) = \sin^2 2\theta \sin^2 \left(\frac{\Delta m^2 \, L}{4E} \right) \quad . \tag{18}$$

Furthermore, as predicted, the survival probability, or disappearance probability is equal to one minus the chance that the neutrino will change from the flavor it was generated with.

3 Neutrino oscillations in a medium

A neutrino beam is created on Earth using an accelerator and sent thousands of kilometers to a detector; instead of traveling through vacuum, the beam travels through matter, or earthly matter. The neutrino beam then disperses from the particles it encounters on its journey and undergo a coherent forward scattering that can (and will) significantly impact the transition probabilities. For the time being, we will make the assumption that neutrino interactions with matter follow the Standard Model's description of flavor conservation and then discuss the potential of flavor-changing interactions. Then, there would only be two scenarios for this coherent forward scattering from matter particles to occur, since there are only two forms of weak interactions (mediated by charged and neutral currents). Weak interactions mediated by charged currents will only happen in the event that the neutrino entering the system is an electron neutrino, since only the ν_e is able to trade a charged boson W for an electron from Earth.

The W exchange mode of neutrino-electron coherent forward scattering thus provides an additional source of interaction energy V_W that is solely experienced by electron neutrinos. Given that it originates from weak interactions, the additional energy must obviously be proportional to the Fermi coupling constant, G_F. Furthermore, the interaction energy resulting from $\nu_e - e$ scattering increases in relation to the number of targets, N_e, which is the electron density per unit volume. When everything is considered, it is evident that

$$V_W = +\sqrt{2}\, G_F \, N_e \quad , \tag{19}$$

obviously, antineutrinos are likewise impacted by this interaction energy (although in the reverse way). If we swap ν_e with $\overline{\nu_e}$, the sign of the interaction energy changes.

Neutral current-mediated interactions are analogous to the situation in which a matter-interacting neutrino exchanges a neutral Z boson with an electron, proton, or neutron. The Standard Model states that weak interactions are independent of flavor. They are enjoyed by all flavors of neutrinos, and the Z exchange's amplitude is constant. It also informs us that protons and electrons bind to the Z boson with identical strength at zero momentum transfer. Nevertheless, the interaction has the opposite sign. Consequently, the contribution of protons and electrons to coherent forward neutrino scattering by Z exchange will add up to zero, presuming that the matter our neutrino travels through is electrically neutral (containing an equal number of protons and electrons). Thus, the influence of the Z exchange contribution to the interaction potential energy V_Z reduces exclusively to that with neutrons and will be proportional to N_n, the number density of neutrons. As a result, only interactions with neutrons will survive. It should go without saying that every flavor will be equal. In such a case we have,

$$V_Z = -\frac{\sqrt{2}}{2}\, G_F\, N_n \ \ , \tag{20}$$

similar to the last instance for V_W, this contribution will reverse if anti-neutrinos are used in place of neutrinos.

However, if, as we previously stated, the Standard Model interactions do not alter the flavor of neutrinos, then even in the case of neutrinos moving through matter, flavor transitions or oscillations clearly indicate the mass and mixing of neutrinos. Unless flavor-changing interactions that aren't standard model-related come into play.

Analyzing neutrino propagation in matter using the lab frame's time-dependent Schrödinger equation makes it simple to understand.

$$i\frac{\partial}{\partial t}|\nu(t)> = \mathcal{H}|\nu(t)> \quad . \tag{21}$$

wherein each neutrino flavor in the (three component) neutrino vector state $|\nu(t)>$ corresponds to a single component. Similarly, in flavor space, the Hamiltonian $\mathcal{H}$ is a (tree × three) matrix. In order to simplify our analysis, let us consider the scenario in which ν_e and ν_μ, the only two neutrino flavors that are important. So that

$$|\nu(t)> = \begin{pmatrix} f_e(t) \\ f_\mu(t) \end{pmatrix} \quad , \tag{22}$$

The neutrino's amplitude at time t to be in, ν_i given by $\mid f_i(t) \mid^2$. This time, $\mathcal{H}$, the Hamiltonian, is a 2×2 matrix in $\nu_e - \nu_\mu$ space, which is the neutrino flavor space.

Working through the two flavor cases in vacuum first and adding matter effects later will prove to be illuminating. For the Hamiltonian in vacuum, $\mathcal{H}_{\text{Vac}}$, we can write $|\nu_\alpha>$ as a linear combination of mass eigenstates using Eq. (2). This allows us to observe that the $\nu_\alpha - \nu_\beta$ matrix element may be expressed as,

$$<\nu_\alpha|\mathcal{H}_{\text{Vac}}|\nu_\beta> = <\sum_i U^*_{\alpha i}\nu_i|\mathcal{H}_{\text{Vac}}|\sum_j U^*_{\beta j}\nu_j> = \sum_j U_{\alpha j}U^*_{\beta j}\sqrt{p^2+m_j^2} \ . \tag{23}$$

where we assume that the neutrinos are part of a beam with the same definite momentum p shared by all of its mass components (the mass eigenstates). As we have already indicated, even if this hypothesis is incorrect in theory, it nevertheless results in the appropriate transition amplitude. The neutrinos ν_j with momentum p, the mass eigenstates, are the asymptotic states of the Hamiltonian, $\mathcal{H}_{\text{Vac}}$ which provide an orthonormal basis, and are employed in the second line of Eq. (23),

$$\mathcal{H}_{\text{Vac}}|\nu_j> = E_j|\nu_j> \tag{24}$$

and the canonical dispersion relation holds, $E_j = \sqrt{p^2 + m_j^2}$.

Neutrino oscillations, as we have already discussed, are the classic example of a quantum interference phenomena, where only the *relative* phases of the interfering states are involved. Consequently, the only values that matter are the *relative* energies of these states, which determine their relative phases. Therefore, we can gladly exclude any contribution proportionate to the identity matrix I from the Hamiltonian $\mathcal{H}$, if that turns out to be useful (which it will). As previously stated, the subtraction will not impact the variations among the eigenvalues of $\mathcal{H}$, and hence, it will not impact the estimation of $\mathcal{H}$ about flavor transitions. Naturally, since only two neutrinos are relevant in this instance, there are only two mass eigenstates, ν_1 and ν_2, and one mass splitting, $\Delta m^2 \equiv m_2^2 - m_1^2$. As a result, a unitary U matrix, given by Eq. (17), should exist as before to rotate from one basis to the other. After inserting it into Eq. (23), assuming that our neutrinos have low masses relative to their momenta, i.e. $(p^2+m_j^2)^{1/2} \cong p+m_j^2/2p$, and eliminating a term proportional to the identity matrix from $\mathcal{H}_{\text{Vac}}$ (a removal we know will be harmless), we obtain

$$\mathcal{H}_{\text{Vac}} = \frac{\Delta m^2}{4E} \begin{pmatrix} -\cos 2\theta & \sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{pmatrix} . \tag{25}$$

The ultra relativistic approximation, which states that $p \cong E$, is utilized to create this expression. where E is the mean energy of the ultra-high momentum p of the neutrino mass eigenstates in our neutrino beam.

The fact that this Hamiltonian $\mathcal{H}_{\text{Vac}}$ of Eq. (25) for the two neutrino scenario would yield the same oscillation probability, Eq. (18), as the one we have previously gotten differently, is easily corroborated. Analyzing the transition probability for the process $\nu_e \longrightarrow \nu_\mu$ is a simple method, for example. It is evident from Eq. (17) that the state describing the electron and muon neutrino is

$$|\nu_e> = \quad |\nu_1> \cos\theta + |\nu_2> \sin\theta \quad , \quad |\nu_\mu> = -|\nu_1> \sin\theta + |\nu_2> \cos\theta \ . \tag{26}$$

The eigenvalues of the vacuum Hamiltonian $\mathcal{H}_{\text{Vac}}$, Eq.25, can also be expressed in terms of the mass squared differences in this manner.

$$\lambda_1 = -\frac{\Delta m^2}{4E} \ , \ \lambda_2 = +\frac{\Delta m^2}{4E} \ . \tag{27}$$

Using Eqs. (26), the mass eigenbasis of this Hamiltonian, $|\nu_1>$ and $|\nu_2>$, may also be expressed in terms of the flavor eigenstates, $|\nu_e>$ and $|\nu_\mu>$. Consequently, the Schrödinger equation of Eq. (21), which identifies $\mathcal{H}$ in this instance with $\mathcal{H}_{\text{Vac}}$, indicates that if we start at a $|\nu_e>$ at time $t = 0$, then after a certain amount of time t passes, this $|\nu_e>$ will advance into the state provided by

$$|\nu(t)> = |\nu_1> e^{+i\frac{\Delta m^2}{4E}t} \cos\theta + |\nu_2> e^{-i\frac{\Delta m^2}{4E}t} \sin\theta \ . \tag{28}$$

Accordingly, from Eqs. (26) and (28), the probability $\text{P}(\nu_e \longrightarrow \nu_\mu)$ that this evolved neutrino be discovered as a new flavor ν_μ is provided by,

$$\text{P}(\nu_e \longrightarrow \nu_\mu) \quad = \quad |<\nu_\mu|\nu(t)>|^2 = |\sin\theta\cos\theta(-e^{i\frac{\Delta m^2}{4E}t} + e^{-i\frac{\Delta m^2}{4E}t})|^2$$

$$= \quad \sin^2 2\theta \sin^2 \left(\Delta m^2 \frac{L}{4E} \right) \quad . \tag{29}$$

Here, the distance L that our extremely relativistic state has traveled is used to replace the time t that it has traveled. As anticipated, the flavor transition or oscillation probability of Eq. (29) is precisely the same as the previous value we obtained, Eq. (18).

The analysis of neutrino propagation in matter can now be continued. In this instance, the two previously stated extra contributions are included into the 2×2 Hamiltonian that represents the propagation in vacuum, $\mathcal{H}_{\text{Vac}}$, to provide $\mathcal{H}_M$, which is given by

$$\mathcal{H}_M = \mathcal{H}_{\text{Vac}} + V_W \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + V_Z \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad . \tag{30}$$

The interaction potential resulting from the exchange of charged bosons is represented by the first additional contribution in the new Hamiltonian, Eq. (19). This contribution differs from zero only in the $\mathcal{H}_M(1{,}1)$ element or the $\nu_e - \nu_e$ element, since this interaction only affects ν_e. The Z boson exchange, Eq. (20), provides the second extra contribution, which is the final component of Eq. (30). It is safe to ignore this interaction as its contribution to $\mathcal{H}_M$ is proportional to the identity matrix, and it impacts all neutrino flavors in the same way because it is flavor blind. Consequently,

$$\mathcal{H}_M = \mathcal{H}_{\text{Vac}} + \frac{V_W}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{V_W}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad , \tag{31}$$

wherein we have split the W-exchange contribution into two parts—one proportionate to the identity (which we will ignore in the following step) and the other, which is not proportional to the identity and which we will retain—for reasons that will become apparent later. As stated, we can disregard the first portion and use the results from Eqs. (25) and (31).

$$\mathcal{H}_M = \frac{\Delta m^2}{4E} \begin{pmatrix} -(\cos 2\theta - A) & \sin 2\theta \\ \sin 2\theta & (\cos 2\theta - A) \end{pmatrix} \quad , \tag{32}$$

where

$$A \equiv \frac{V_W/2}{\Delta m^2/4E} = \frac{2\sqrt{2} G_F N_e E}{\Delta m^2} \quad . \tag{33}$$

The relative magnitude of the matter effects in relation to the vacuum contribution provided by the neutrino squared-mass splitting is clearly parameterized by A, which also indicates the circumstances in which these effects become significant.

Now, if we introduce a shorthand notation of physical significance

$$\Delta m_M^2 \equiv \Delta m^2 \sqrt{\sin^2 2\theta + (\cos 2\theta - A)^2} \tag{34}$$

and

$$\sin^2 2\theta^M \equiv \frac{\sin^2 2\theta}{\sin^2 2\theta + (\cos 2\theta - A)^2} \quad , \tag{35}$$

then the Hamiltonian describing the propagation of the neutrinos in matter $\mathcal{H}_M$ becomes

$$\mathcal{H}_M = \frac{\Delta m_M^2}{4E} \begin{pmatrix} -\cos 2\theta^M & \sin 2\theta^M \\ \sin 2\theta^M & \cos 2\theta^M \end{pmatrix} . \tag{36}$$

The Hamiltonian in a medium, $\mathcal{H}_M$, becomes formally identical from the vacuum one, $\mathcal{H}_{\text{Vac}}$, Eq. (25), as a result of our definitions. Then can be trivially diagonalized. The difference between the Hamiltonians resides in the fact that, the matter parameters, Δm_M^2 and θ^M, respectively, become now, what were previously the vacuum parameters, Δm^2 and θ.

It is evident that the mass eigenstates and eigenvalues of $\mathcal{H}_M$ differ from those in vacuum, but are still given by the mass differences and mixing angle in matter in the same way as the ones in vacuum before. Thus, the vacuum eigenvalues that make up the vacuum mixing matrix are distinct from the eigenstates in matter, i.e. the values of the unitary matrix that rotates from the flavor basis to the mass basis are not the same and obviously, θ_M is not equal to θ. However, just as $\mathcal{H}_{\text{Vac}}$ contains all the information regarding neutrino propagation in vacuum, the matter Hamiltonian $\mathcal{H}_M$ does in fact contain all the information regarding neutrino propagation in matter.

The functional dependence of $\mathcal{H}_M$ on the matter parameters Δm_M^2 and θ^M is the same as that of the vacuum Hamiltonian $\mathcal{H}_{\text{Vac}}$, Eq. (25), on the vacuum ones, Δm^2 and θ, according to Eq. (36). As a result, θ^M can be identified with an effective mixing angle in matter, and Δm_M^2 can be identified with an effective mass squared difference in matter.

In a standard experimental setup, the neutrino beam travels through earth matter just superficiall—it does not penetrate deeply—after being produced by an accelerator and directed toward a detector located hundreds or even thousands of kilometers away. Therefore, it may be assumed that the matter density this beam would meet is roughly constant during its journey. However, it is evident that this approximation is invalid for neutrinos traveling over the Earth. However, this also holds true for the electron density N_e and the A parameter in which it is included if the density of matter on Earth remains constant. Regarding the Hamiltonian $\mathcal{H}_M$, it is also accurate. Except for their specific parameter values, that as we have seen can be considered all almost constant, the situation is identical to the one we faced with the vacuum Hamiltonian $\mathcal{H}_{\text{Vac}}$. Therefore, we already know that, in the same way that $\mathcal{H}_{\text{Vac}}$ gives rise to vacuum oscillations with probability $\mathrm{P}(\nu_e \longrightarrow \nu_\mu)$ of Eq. (29), $\mathcal{H}_M$ must give rise to matter oscillations, which by comparing Eqs. (36) and (25), is given by

$$\mathrm{P}_M(\nu_e \longrightarrow \nu_\mu) = \sin^2 2\theta^M \sin^2\left(\Delta m_M^2 \frac{L}{4E}\right) . \tag{37}$$

In other words, the survival and transition probabilities in matter are the same as in vacuum, with the exception that the parameters in vacuum, Δm^2 and θ, are now substituted by the corresponding values in matter, Δm_M^2 and θ^M.

Merely based on its potential, matter effects have the ability to significantly alter oscillation probabilities, at least in theory. Only until the specifics of the experiment's experimental setup are provided can the precise impact, if any, be calculated. Generally speaking, if the kinematic phase linked to the solar mass difference is still insignificant and neutrinos are traveling through the earth's mantle (no more than 200 km below the surface), then studying

$$A \cong \frac{E}{13 \ \text{GeV}} \tag{38}$$

will help us estimate the significance of matter effects. And we can easily see that matter effects are relevant only for beam intensities of few GeV.

And what effect do they have? They are really important! We can observe from Eq. (35) for the matter mixing angle, θ^M, that even in cases when the vacuum mixing angle θ is minuscule, for example, $\sin^2 2\theta = 10^{-4}$, in comparison to its vacuum value, $\sin^2 2\theta^M$ can be greatly boosted if we can obtain $A \cong \cos 2\theta$, i.e., for energies of a few tens of GeV. It can even reach maximal mixing, $\sin^2 2\theta^M = 1$.

This amazing enhancement, known as the Mikheyev-Smirnov-Wolfenstein effect [5, 6], is a "resonant" amplification of a small mixing angle in vacuum up to a significant one in matter, up to maximal. When solar neutrino physics first started, there was a theory that this violent amplification was really happening when neutrinos traveled across the sun. However, as we shortly afterward observed, the mixing angle linked to solar neutrinos is already relatively large ($\sim 34°$) in vacuum [7]. Consequently, while matter effects on the sun are indeed significant, they are sadly not as significant as we had originally anticipated. However, over long-baselines, they will—and already do—play a crucial part in deciding the ordering of the neutrino mass spectrum.

4 Evidence for neutrino oscillations

4.1 Atmospheric and accelerator neutrinos

Since we were first shown strong, persuasive evidence of neutrino masses and mixings more than two decades ago, the body of evidence has only increased. The first experiment to provide strong evidence of ν_μ disappearance in their atmospheric neutrino fluxes was SuperKamiokande (SK) [8]. The multi-GeV ν_μ sample's zenith angle dependency, or the angle subtended with the horizontal, and its disappearance as a function of L/E are displayed in Fig. 2. These findings fit the simple two-component neutrino theory with remarkable accuracy with

$$\Delta m^2_{\text{atm}} = 2 - 3 \times 10^{-3} \text{eV}^2 \quad \text{and} \quad \sin^2 \theta_{\text{atm}} = 0.50 \pm 0.13 \tag{39}$$

for oscillations of L/E of 500 km/GeV and nearly maximal mixing, suggesting that the mass eigenstates are nearly even admixtures of tau and muon neutrinos. Since the third flavor, ν_e, does not exhibit any evidence of involvement, it is assumed that atmospheric neutrino disappearance is essentially $\nu_\mu \longrightarrow \nu_\tau$. However, take note of the fact that more recent results strongly suggest a mixing angle that is not maximal.

Following the discovery of atmospheric neutrino oscillations, a new set of neutrino experiments was constructed, sending (man-made) beams of ν_μ neutrinos to detectors situated at great distances: the MINOS (NOvA) experiment [9, 10] sends its beam from Fermilab, near Chicago, to the Soudan mine (Ash river) in Minnesota, a baseline of 735 (810) km, while the K2K (T2K) experiment [11, 12] sends neutrinos from the KEK accelerator complex to the old SK mine, with a baseline of 120 (235) km. Evidence of ν_μ disappearance consistent with the one found by SK has been observed in all these

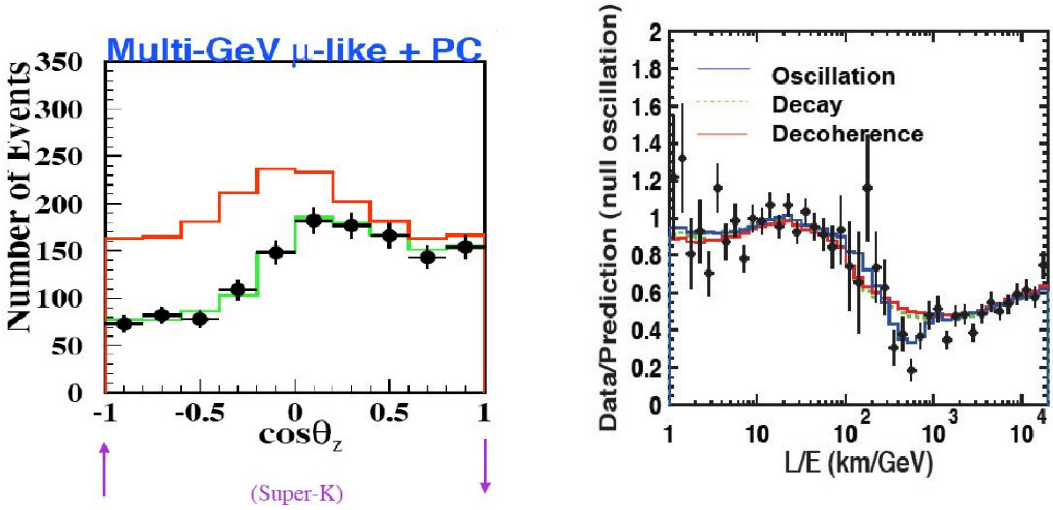


Fig. 2: Superkamiokande's evidence for neutrino oscillations both in the zenith angle and L/E plots.

studies. Figure 3 summarizes their findings.

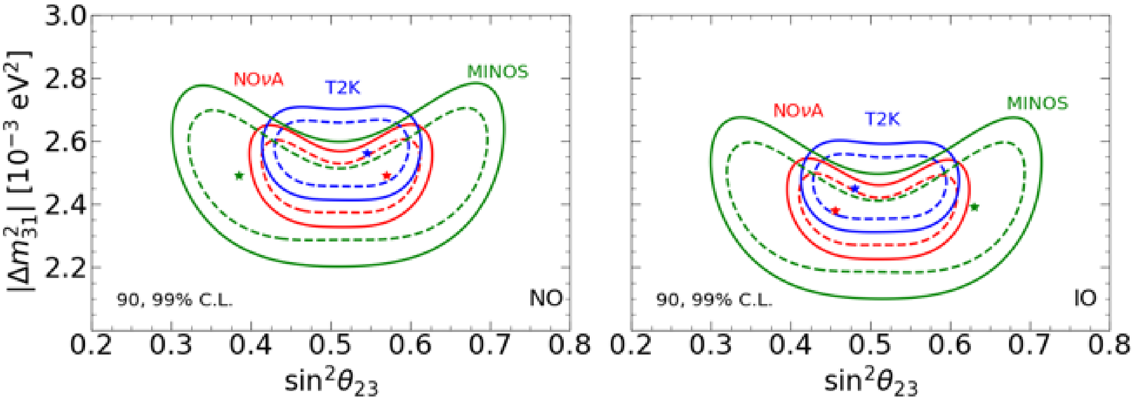


Fig. 3: Allowed regions in the $\Delta m^2_{\rm atm}$ vs $\sin^2\theta_{\rm atm}$ plane for MINOS and NOVA data as well as for T2K data and two of the SK analyses. Results from https://globalfit.astroparticles.es/

4.2 Reactor and solar neutrinos

Evidence of neutrino oscillations has been observed in the KamLAND reactor experiment, an antineutrino disappearance experiment that receives neutrinos from sixteen different reactors at distances ranging from hundreds to thousands of kilometers, with an average baseline of 180 km and neutrinos of a few eV [13]. This proof was gathered not only at a distinct L/E compared to the atmosphere and accelerator studies, but also includes oscillations involving electron neutrinos, ν_e, which were not implicated before. Since the sun only creates electron neutrinos, these oscillations have also been observed in neutrinos that originate from it. However, we must make the assumption that neutrinos (from the sun) and antineutrinos (from the reactor) behave similarly in order to compare the two tests; this is known as CPT conservation.

For the KamLAND experiment, the best fit values in the two neutrino scenario are

$$\Delta m^2_\odot = 7.55 \pm 0.2 \times 10^{-5}\text{eV}^2 \quad \text{and} \quad \sin^2\theta_\odot = 0.32 \pm 0.03 \tag{40}$$

The mixing angle, while high, is obviously not maximal in this situation, and the L/E involved is 15 km/MeV.

The $\bar{\nu}_e$ disappearance probability for KamLAND and a few older reactor studies with shorter baselines are displayed in Fig. 4 as well as the best fit regions. The analysis of neutrinos originating from the sun is slightly more complex than the previous one we conducted. This is especially true for the ^{8}Boron neutrinos. Due to their lower energy, the pp and ^{7}Be neutrinos are not greatly affected by the existence of matter. This is not the case for ^{8}Boron neutrinos which depart the sun as ν_2, the second mass eigenstate, and do not oscillate, clearly indicating that they are affected by the existence of matter.

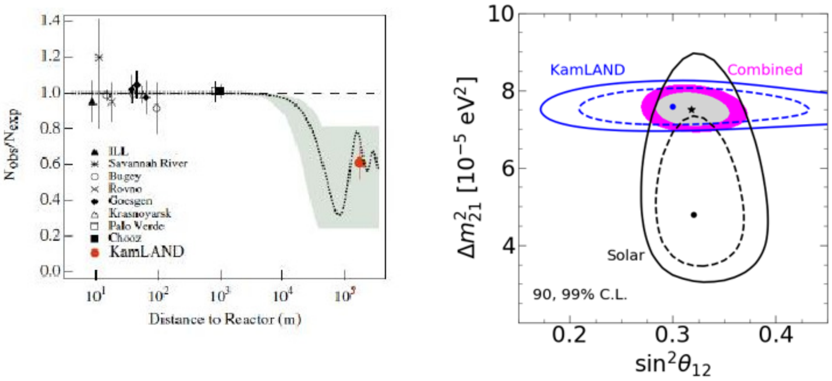


Fig. 4: Disappearance of the $\bar{\nu}_e$ observed by reactor experiments as a function of distance from the reactor. Favored region for all solar and reactor experiments. Results from https://globalfit.astroparticles.es/

It is important to note, nevertheless, that solar neutrino oscillations are not really observed. We require a kinematic phase of order one in order to follow the oscillation pattern and test its unique shape; otherwise, the oscillations either do not emerge or average out to 1/2. For neutrinos that originate from the sun, the kinematic phase is

$$\Delta_\odot = \frac{\Delta m^2_\odot L}{4E} = 10^{7\pm1} \quad . \tag{41}$$

As a result, when solar neutrinos leave the sun, they behave as "effectively incoherent" mass eigenstates, and they stay that way when they get to Earth. As a result, the ν_e survival or disappearance probability is provided by

$$\langle P_{ee}\rangle = f_1 \cos^2\theta_\odot + f_2 \sin^2\theta_\odot \tag{42}$$

where f_1 (f_2) is the $\nu_1(\nu_2)$ content of ν_μ, and $f_1 + f_2 = 1$.

In the case of solar neutrinos originating from the pp and ^{7}Be chains, the fractions are $f_1 \approx \cos^2\theta_\odot = 0.69$ and $f_2 \approx \sin^2\theta_\odot = 0.31$ since they oscillate as in vacuum and are unaffected by solar matter. However, the impact of solar matter is significant, and the corresponding fractions are significantly changed in the ^{8}B neutrino case in which

$$f_2 = 0.91 \pm 0.02 \text{ at the 95 \% C.L. } . \tag{43}$$

As a result, it is evident that the ^{8}B solar neutrinos are the cleanest mass eigenstate neutrino beam currently known.

Last but not least, the third mixing angle [14] was finally measured eleven years ago by the Daya Bay experiment, a reactor neutrino experiment in China. It was discovered to be

$$\sin^2(2\theta_{13}) = 0.092 \pm 0.017 \ . \tag{44}$$

Subsequent to this discovery, other experiments validated the discovery, and in recent years, the last mixing angle measured emerged as the most accurate one. This angle's sizeable size, albeit being smaller than the other two, allows for the possibility of future neutrino studies that seek to address the unanswered concerns in the field.

5 ν Standard Model

After understanding the physics underlying neutrino oscillations and gaining insight from experimental data regarding the parameters causing these oscillations, we can proceed to build the Neutrino Standard Model, which consists of three light ($m_i < 1$ eV) neutrinos, or i.e. only two mass differences.

As of yet, there has been no conclusive, or even solid, experimental evidence supporting the need for more neutrinos, though it should be emphasized that there are a few weak signals. Since the invisible width of the Z boson has been measured for a long time and is found to be 3, within errors, any more neutrinos added to the model will have to be sterile since they will not be able to couple to the Z boson, i.e. they will not be able to experience weak interactions. Nevertheless, our Neutrino Standard Model will only include the three active flavors, e, μ, and τ, as sterile neutrinos have not been observed and are not required to explain any experimental data.

Three mixing angles, the so-called solar mixing angle: θ_{12}, the atmospheric mixing angle: θ_{23}, and the final one to be measured, the reactor mixing angle: θ_{13}, one Dirac phase, δ, and possibly two Majorana phases, $\alpha$and β, make up the unitary mixing matrix, also known as the PMNS matrix, which rotates from the flavor to the mass basis. It is given by

$$\mid \nu_\alpha\rangle = U_{\alpha i} \mid \nu_i\rangle \ ,$$

$$U_{\alpha i} = \begin{pmatrix} 1 & & \\ & c_{23} & s_{23} \\ & -s_{23} & c_{23} \end{pmatrix} \begin{pmatrix} c_{13} & & s_{13}e^{-i\delta} \\ & 1 & \\ -s_{13}e^{i\delta} & & c_{13} \end{pmatrix} \begin{pmatrix} c_{12} & s_{12} & \\ -s_{12} & c_{12} & \\ & & 1 \end{pmatrix} \begin{pmatrix} 1 & & \\ & e^{i\alpha} & \\ & & e^{i\beta} \end{pmatrix}$$

where $s_{ij} = \sin\theta_{ij}$ and $c_{ij} = \cos\theta_{ij}$. We can identify the (23) label in the three neutrino scenario as the atmospheric Δm^2_{atm} that we obtained in the two neutrino scenario thanks to the hierarchy in mass differences (and, to a lesser extent, the smallness of the reactor mixing angle). Similarly, the (12) label can be assimilated to the solar $\Delta m^2_{\odot}$. The ν_e flavor oscillations at the atmospheric scale are driven by the (13) sector.

As per current experimental results, the three sigma ranges associated with the neutrino parameters are

$0.271 < \sin^2\theta_{12} < 0.369$; $0.434 < \sin^2\theta_{23} < 0.610$; $0.0200 < \sin^2\theta_{13} < 0.0245$ $2.47 \times 10^{-3}\mathrm{eV}^2 < |\Delta m^2_{32}| < 2.63 \times 10^{-3}\mathrm{eV}^2$ and $6.94 \times 10^{-5}\mathrm{eV}^2 < \Delta m^2_{21} < 8.14 \times 10^{-5}\mathrm{eV}^2$.

Two orderings are possible because oscillation experiments only investigate the two mass differences. They are known as normal and inverted hierarchy, and indicate whether the lightest or heaviest mass eigenstate, respectively, is the one with the smallest electron neutrino content.

Although the mass of the lightest neutrino, or the absolute mass scale of neutrinos, is unknown, cosmological bounds already dictate that the heaviest neutrino must be lighter than roughly 2 eV.

No evidence of the Majorana phases could be seen in oscillation phenomena because transition or survival probabilities rely on the combination $U^*_{\alpha i}U_{\beta i}$. However, they will be noticeable in processes where the Majorana character of the neutrino is necessary for the process to occur, such as neutrino-less double beta decay.

6 Neutrino mass and character

6.1 Absolute neutrino mass

Although oscillation experiments are unable to yield the neutrino's absolute mass scale, or the mass of the lightest/heaviest neutrino, this does not mean that we are without means of obtaining it. Both direct experiments—such as tritium beta decay or neutrinoless double beta decay—and indirect ones—such as cosmic observations—have the capacity to provide us with the much-needed details on the precise scale of neutrino mass. The sensitivity of the Katrin tritium beta decay experiment, [15], given the "mass" of ν_e defined as

$$m_{\nu_e} = | U_{e1} |^2 m_1 + | U_{e2} |^2 m_2 + | U_{e3} |^2 m_3. \tag{45}$$

Neutrino-less double beta decay experiments assess a specific mixture of neutrino masses and mixings rather than the neutrino's absolute mass, see for example [16].

$$m_{\beta\beta} = | \sum m_i U^2_{ei} | = | m_a c^2_{13} c^2_{12} + m_2 c^2_{13} s^2_{12} e^{2i\alpha} + m_3 s^2_{13} e^{2i\beta} |, \tag{46}$$

wherein it is assumed that neutrinos are Majorana particles—that is, completely neutral particles. particles with zero quantum numbers throughout. In double beta decay, the goal of the latest studies is to lower the energy of $m_{\beta\beta}$ below 10 meV.

The total mass of neutrinos is measured by cosmological experiments [17] such as the Large Scale Structure experiments and the CMB. It is defined as

$$m_{\text{cosmo}} = \sum_i m_i \tag{47}$$

and allows to investigate additional features of neutrinos, such as neutrino asymmetries, and the mass ordering [18]. The present limit is ~ 0.2 eV. Although these bounds depend on the model, they all produce numbers with the same order of magnitude. As cosmological measurements are characterized by systematic errors, a definite limit of less than 200 meV appears much too aggressive.

6.2 Majorana vs Dirac

A coupling between a left-handed and a right-handed state is all that a fermion mass is. As a result, we can think of a massive fermion at rest as a linear combination of two massless particles, one left-handed and one right-handed. Both the left and right-handed particles must have the same charge if the particle under study is electrically charged, such as an electron or a muon[1]. We have then a Dirac mass.

Thus, a completely and totally neutral particle, which is inevitable to become its own antiparticle, has two options for obtaining a mass term: Majorana or Dirac. If neither option is prohibited, it will have both.

In the case of a neutrino, the left chiral field couples to $SU(2) \times U(1)$ indicating that a Majorana mass term is forbidden by gauge symmetry. However, the right chiral field carries no quantum numbers and is totally and absolutely neutral. Then, the Majorana mass term is unprotected by any symmetry, and it is expected to be very large, of the order of the largest scale in the theory. On the other hand, Dirac mass terms are expected to be of the order of the electroweak scale times a Yukawa coupling, giving a mass of the order of magnitude of the charged lepton or quark masses. Putting all the pieces together, the mass matrix for the neutrinos results in

$$\begin{pmatrix} 0 & m \\ m & M \end{pmatrix} .$$

The neutrino mass matrix must be diagonalized in order to obtain the mass eigenstates. When this is done, one is left with two Majorana neutrinos: one super-light Majorana neutrino with mass m_D^2/M and one super-heavy Majorana neutrino with mass $\simeq M$. This is known as the seesaw mechanism because one mass rises while the other falls [19–21]. The simplest see-saw mechanisms can be divided into three groups based on their scalar content, depending on the high energy theory that is envisaged. While the heavy neutrinos are not accessible to current experiments, they may be responsible for explaining the baryon asymmetry of the universe by generating a lepton asymmetry at very high energy scales. This is because their decays have the potential to be CP violating because they depend on the two Majorana phases on the PNMS matrix, which are invisible for oscillations. The light neutrinos are those that are observed in current experiments. Due to their enormous masses, the very heavy Majorana neutrinos may contribute to inflation at very high energy [22].

Lepton number is no longer a valid quantum number if neutrinos are Majorana particles, and a variety of novel processes that are prohibited by lepton number conservation can occur—not just neutrino-less double beta decay. For instance, a positively charged muon can be created by a muon neutrino. Though mathematically permissible, these processes—and all others of their kind—would be suppressed by the minuscule 10^{-20} of $(m_\nu/E)^2$. As a result, they are experimentally unobservable. At

[1] We want the mass term to be electrically neutral

90% C.L., the most rigorous limit currently available for neutrino-less double beta decay comes from KamLAND-zen [23], which limits the half-life to $T_{1/2}^{0\nu} > 2.3 \times 10^{26}$ years. This sensitivity will increase by a factor of ten in the not so distant future.

Lately, low energy sew saw models [24] have come back into vogue and are being actively studied [25]. The LHC can be used to hunt for heavy states in these models that have energies of only a few tens of TeV. In these theories, the heavy right-handed states will be generated at the LHC by gauge coupling to right-handed gauge bosons or via Yukawa couplings.

7 Conclusions

After the Standard Model was established, there were still unanswered questions. These were resolved by the experimental observations of neutrino oscillations, which indicates that neutrinos have mass and mix. With the removal of those coverings, fresh inquiries arise that cast doubt on our comprehension.

What are neutrinos actually, Dirac or Majorana particles? Does a new scale related to the mass of neutrinos exist? Is it available at colliders? Is the spectrum inverted or normal? Which neutrino is lighter, the one with the fewest electron content on it, or is it the heaviest one? Is $\sin\delta \neq 0$? If yes, how is this phase connected to the universe's baryon asymmetry at all? Which is the neutrino's absolute mass scale? Do unanswered cosmological concerns like dark energy and/or dark matter have anything to do with neutrinos? Are (supposedly massive) neutrinos involved in the primordial inflation? Is CPT violated by neutrinos [26]? Lorentz invariance: what about it? How can we determine whether a non-standard neutrino interaction or a true CTP violation is the cause of a difference in the spectrum measured for neutrinos and antineutrinos if we ever measure it?

We would want to respond to these questions. We are already doing it, and we intend to do other studies. These experiments will undoubtedly raise important new questions and provide some solutions. There's only one thing that's obvious. Our exploration of the neutrino universe is only getting started.

Acknowledgements

I am grateful to the students and the organizers of the European School on HEP for providing me with this amazing platform to deliver these lectures. I thoroughly enjoyed every day of school. This work was supported by the Spanish grants PID2020-113775GB-I00 (AEI/10.13039/501100011033) and CIPROM/2021/054 (Generalitat Valenciana)

References

[1] *Symmetry* **4** (2007) ibc,
https://www.symmetrymagazine.org/article/march-2007/neutrino-invention.

[2] Z. Maki, M. Nakagawa and S. Sakata, *Prog. Theor. Phys.* **28**, (1962) 870–880,
doi:10.1143/PTP.28.870.

[3] L. Stodolsky, *Phys. Rev.* **D58** (1998) 036006, doi:10.1103/PhysRevD.58.036006,
arXiv:hep-ph/9802387; H.J. Lipkin, *Phys. Lett.* **B579** (2004) 355–360,
doi:10.1016/j.physletb.2003.11.013, arXiv:hep-ph/0304187.

[4] An analysis of CPT violation in the neutrino sector can be found in: G. Barenboim and J.D. Lykken, *Phys. Lett.* **B554** (2003) 73–80, doi:10.1016/S0370-2693(02)03262-8, arXiv:hep-ph/0210411; G. Barenboim, J.F. Beacom, L. Borissov and B. Kayser, *Phys. Lett.* **B537** (2002) 227–232, doi:10.1016/S0370-2693(02)01947-0, arXiv:hep-ph/0203261; The world best bound in CPT conservation is found in G. Barenboim, C.A. Ternes and M. Tortola, *Phys. Lett.* **B780** (2018) 631–637, doi:10.1016/j.physletb.2018.03.060, arXiv:1712.01714 [hep-ph].

[5] L. Wolfenstein, *Phys. Rev.* **D17** (1978) 2369–2374, doi:10.1103/PhysRevD.17.2369.

[6] S.P. Mikheev and A.Y. Smirnov, *Sov. J. Nucl. Phys.* **42** (1985) 913 [*Yad. Fiz.* **42** (1985) 1441]; S.P. Mikheev and A.Y. Smirnov, *Sov. Phys. JETP* **64** (1986) 4 [*Zh. Eksp. Teor. Fiz.* **91** (1986) 7, arXiv:0706.0454 [hep-ph]; S.P. Mikheev and A.Y. Smirnov, *Nuovo Cim.* **C9** (1986) 17, doi:10.1007/BF02508049.

[7] N. Tolich [SNO Collaboration], *J. Phys. Conf. Ser.* **375** (2012) 042049, doi:10.1088/1742-6596/375/1/042049.

[8] V. Takhistov [Super-Kamiokande], *PoS* **ICHEP2020** (2021), 181, doi:10.22323/1.390.0181.

[9] J. Evans [MINOS and MINOS+], *J. Phys. Conf. Ser.* **888** (2017) 012017, doi:10.1088/1742-6596/888/1/012017.

[10] J.M. Carceller [NOvA], *PoS* **NOW2022** (2023) 015, doi:10.22323/1.421.0015.

[11] M.H. Ahn *et al.* [K2K], *Phys. Rev.,***D74** (2006) 072003, doi:10.1103/PhysRevD.74.072003, arXiv:hep-ex/0606032.

[12] L. Berns [T2K], *PoS* **NOW2022** (2023) 002, doi:10.22323/1.421.0002.

[13] M.P. Decowski [KamLAND], *Nucl. Phys.* **B908** (2016) 52–61, doi:10.1016/j.nuclphysb.2016.04.014.

[14] F.P. An *et al.* [Daya Bay], *Phys. Rev. Lett.* **129** (2022) 041801, doi:10.1103/PhysRevLett.129.041801.

[15] S. Mertens [KATRIN Collaboration], *Phys. Procedia* **61** (2015) 267, doi:10.1016/j.phpro.2014.12.043.

[16] S.R. Elliott and P. Vogel, *Ann. Rev. Nucl. Part. Sci.* **52** (2002) 115, doi:10.1146/annurev.nucl.52.050102.090641, arXiv:hep-ph/0202264.

[17] M. Lattanzi [Planck Collaboration], *J. Phys. Conf. Ser.* **718** (2016) 032008, doi:10.1088/1742-6596/718/3/032008.

[18] G. Barenboim, W.H. Kinney and W.I. Park, *Phys. Rev.* **D95** (2017) 043506, doi:10.1103/PhysRevD.95.043506, arXiv:1609.01584 [hep-ph]; G. Barenboim, W.H. Kinney and W.I. Park, *Eur. Phys. J.* **C77** (2017) 590, doi:10.1140/epjc/s10052-017-5147-4.

[19] M. Gell-Mann, P. Ramond and R. Slansky, in *Supergravity*, edited by P.van Nieuwenhuizen and D. Freedman, (North-Holland,1979), p.315.

[20] R. N. Mohapatra and G. Senjanovic, Phys. Rev. Lett. **44**, 912 (1980).

[21] M. Fukugita and T. Yanagida, Phys. Lett. B **174**, 45 (1986).

[22] G. Barenboim, JHEP **0903** (2009) 102 [arXiv:0811.2998]; G. Barenboim, Phys. Rev. D **82** (2010) 093014 [arXiv:1009.2504].

[23] K. Ichimura [KamLAND-Zen], PoS **NOW2022** (2023), 067 doi:10.22323/1.421.0067

[24] F. Borzumati and Y. Nomura, Phys. Rev. D **64** (2001) 053005 doi:10.1103/PhysRevD.64.053005 [hep-ph/0007018].

[25] C. G. Cely, A. Ibarra, E. Molinaro and S. T. Petcov, Phys. Lett. B **718** (2013) 957 doi:10.1016/j.physletb.2012.11.026 arXiv[1208.3654].

[26] G. Barenboim, L. Borissov, J. D. Lykken and A. Y. Smirnov, JHEP **0210** (2002) 001 [hep-ph/0108199]. G. Barenboim and J. D. Lykken, Phys. Rev. D **80** (2009) 113008 arXiv[0908.2993].

Flavour physics and CP violation

Alexander Lenz[a]
[a]University of Siegen, Siegen, Germany

We give a brief introduction into quark flavour physics and CP violation, starting in the first lecture with a review of the fundamental properties of the Standard Model of Particle Physics, a detailed discussion of the CKM matrix and a general classification of hadronic weak decays. The second lecture is devoted to describing the theoretical framework and in particular the concept of an effective Hamiltonian. In the third lecture we discuss mixing of neutral mesons and the effect of CP violation in hadron decays.

This chapter should be cited as: Flavour physics and CP violation, Alexander Lenz, DOI: 10.23730/CYRSP-2025-009.145, in: Proceedings of the 2023 European School of High-Energy Physics,
CERN Yellow Reports: School Proceedings, CERN-2025-009, DOI: 10.23730/CYRSP-2025-009, p.145.

1 Lecture 1: Standard Model, CKM matrix, weak decays

1.1 Motivation for Flavour Physics

The field of **Quark Flavour Physics** offers interesting opportunities in deepening our fundamental understanding of the world. In particular it sheds light into:

1. Matter-Antimatter asymmetry in the Universe: as we will see below the existence of matter in the Universe seems to be linked to the breaking of a symmetry, called CP. CP-violating effects were observed numerous times in the decays of hadrons containing heavy quarks.
2. Indirect searches for effects beyond the standard model (SM): finding discrepancies, when comparing precise measurements with precise SM predictions might give us hints how to extend the SM on a more fundamental level. This strategy benefits currently from the huge amount of data obtained by experiments like LHCb, Belle II, BES III, ATLAS and CMS. It is interesting to note that there are currently some anomalies, i.e. deviations from the SM expectations, observed in the decays of b- and c-hadrons.
3. Understanding of QCD: for the program of indirect new physics searches a rigorous control over hadronic effects is crucial. The theoretical description of these processes relies on different variations of effective theories.
4. Determination of SM parameter: the bread and butter physics goal of flavour physics is the determination of SM parameters like CKM-elements or quarks masses.

1.2 Lagrangian of the Standard Model

The Lagrangian of the Standard Model of Particle Physics (SM) [1–3] reads schematically

$$\begin{aligned}\mathcal{L} &= -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \\ &\quad +i\bar{\Psi}\,\not{D}\Psi \\ &\quad +|D_\mu\Phi|^2 - V(\Phi) \\ &\quad +\bar{\Psi}_i Y_{ij}\Phi\Psi_j + h.c. \quad . \end{aligned} \tag{1}$$

The first line of Eq. (1) is the gauge kinetic term, which describes the propagation of the massless gauge fields of the strong, weak and electro-magnetic interactions, as well as, in the case of the latter two, their self-interaction. The second line is the fermionic kinetic term, which describes the propagation of massless fermions and their interactions with the gauge fields. The third line is known as the Higgs sector of the SM and contains the kinetic term for the complex Higgs doublet, including its interaction with the gauge fields, and the Higgs potential, see [4–7]. The specific form of the Higgs potential will give rise to mass terms for some of the gauge bosons. The last line is known as the Yukawa sector, which describes the interaction between the fermion fields Ψ and the complex scalar field Φ. When the Higgs field acquires the vacuum expectation value $(0, v/\sqrt{2})$, a mass term for the fermion fields is generated, with the mass being proportinal to $vY_{ij}/\sqrt{2}$.

The full Standard Model Lagrangian is invariant under Poincare transformations and local $SU(3)_C \times SU(2)_L \times U(1)_Y$ gauge transformations — $SU(3)_C$ describes the strong interaction, C stands for colour,

$SU(2)_L \times U(1)_Y$ describes the weak and the electromagnetic interaction, L stands for left-handed, and Y stands for hyper-charge. Looking at the $SU(2)_L \times U(1)_Y$-part in more detail, one gets in the case of one generation of fermions the following expressions:

$$\begin{aligned}\mathcal{L}_{SU(2)_L \times U(1)_Y} &= -\frac{1}{4} W^a_{\mu\nu} W^{\mu\nu\, a} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} \\ &+ \bar{\Psi}_L \gamma^\mu \left(i\partial_\mu - g_1 Y_L B_\mu - g_2 q_L \frac{\vec{\sigma} \cdot \vec{W}_\mu}{2} \right) \Psi_L \\ &+ \bar{\Psi}_R \gamma^\mu \left(i\partial_\mu - g_1 Y_R B_\mu - g_2 q_R \frac{\vec{\sigma} \cdot \vec{W}_\mu}{2} \right) \Psi_R \\ &+ \left| \left(i\partial_\mu - g_1 Y_\Phi B_\mu - g_2 q_\Phi \frac{\vec{\sigma} \cdot \vec{W}_\mu}{2} \right) \Phi \right|^2 - V(\Phi^\dagger \Phi) \\ &- \left(\bar{\Psi}_L \Phi^c Y_u u_R + \bar{u}_R \Phi^{c\dagger} Y_u^\dagger \Psi_L \right) - \left(\bar{\Psi}_L \Phi Y_d d_R + \bar{d}_R \Phi^\dagger Y_d^\dagger \Psi_L \right) . \end{aligned} \tag{2}$$

Let us first discuss the notation:

a) Ψ_L and Ψ_R denote left- and right-handed spinors describing the fermion fields

$$\Psi_{L,R} = \frac{1 \mp \gamma_5}{2} \Psi . \tag{3}$$

The splitting into left- and right-handed components is motivated by the experimental observation of parity violation in weak decays. The violation of parity in the weak interaction was theoretically proposed in 1956 by Lee and Yang (Nobel Prize 1957) [8] and almost immediately verified in the experiment of Wu [9]. A way of implementing this fact into the theory is treating the right-handed fermions Ψ_R as $SU(2)_L$ singlets and the left-handed fermions Ψ_L as $SU(2)_L$ doublets, i.e.

$$\Psi_L = \begin{pmatrix} u_L \\ d_L \end{pmatrix} . \tag{4}$$

Here, u_L is the four component Dirac spinor field of the up-quark, with weak isospin $+1/2$ and d_L is the four component Dirac spinor field of the down quark with weak isospin $-1/2$.

b) g_1 is the gauge coupling of the $U(1)_Y$ interaction mediated via the gauge field B_μ, with the corresponding field strength tensor defined as

$$B_{\mu\nu} := \partial_\mu B_\nu - \partial_\nu B_\mu . \tag{5}$$

$Y_{L,R,\phi}$ are the hyper-charges of the left-handed fermions, right-handed fermions and of the Higgs field.

c) g_2 is the gauge coupling of the $SU(2)_L$ interaction mediated via the three gauge fields, $W^a_\mu (a = 1, 2, 3)$, with the corresponding field strength tensor, defined as

$$W^a_{\mu\nu} := \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g_2 \epsilon^{abc} W^b_\mu W^c_\nu , \tag{6}$$

where ϵ^{abc} are the structure constants of the $SU(2)$ algebra, i.e.

$$\left[\frac{\sigma^a}{2}, \frac{\sigma^b}{2}\right] = i\epsilon^{abc}\frac{\sigma^c}{2} \tag{7}$$

and $\vec{\sigma} = (\sigma^1, \sigma^2, \sigma^3)$ denote the Pauli matrices. Note the last term in Eq. (6) gives rise to self-interaction among the W^a_μ fields. The fact that only left-handed fermions take part in the weak interaction and right-handed do not, is fulfilled by the following choice of the charges: $q_R = 0$ and $q_L = q_\Phi = 1$. This describes correctly the experimentally found *maximal parity-violation* of the weak interaction.

d) Also the Higgs field is a $SU(2)_L$ doublet

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}, \tag{8}$$

with hyper-charge $Y = 1/2$. The complex Higgs doublet has four degrees of freedom and the quantum numbers,

	ϕ^+	ϕ^0
Q	$+1$	0
T_3	$+1/2$	$-1/2$
Y	$+1/2$	$+1/2$

.

Using the **unitary gauge** the Higgs field is expanded as

$$\Phi = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + H \end{pmatrix}, \tag{9}$$

where $\langle\Phi\rangle = 1/\sqrt{2}(0, v)$ is the non-vanishing vacuum expectation value of the Higgs doublet Φ with $v \approx 246.22$ GeV,[1] and H describes the Higgs particle discovered at the LHC in 2012 [10,11]. To give both the up-type and the down-type quarks a mass we have to introduce a second Higgs field, which is not independent from the original one (in some extensions of the standard model, it will be independent, e.g. in the Two-Higgs Doublet Model (**2HDM**) or the Minimal Supersymmetric Standard Model (**MSSM**)), i.e.

$$\Phi^c = i\sigma_2\Phi^* = \begin{pmatrix} \phi^{0*} \\ -\phi^{+*} \end{pmatrix}, \tag{10}$$

which can also be expanded as

$$\Phi^c = \frac{1}{\sqrt{2}}\begin{pmatrix} v + H \\ 0 \end{pmatrix}. \tag{11}$$

[1] Originally v is defined as the minimum of the Higgs potential, $v = \sqrt{-\mu^2/\lambda}$. Expressing the gauge boson masses in terms of v one gets $M_W = g_2 v/2$. Comparing this with the definition of the Fermi constant $G_F/\sqrt{2} = g_2^2/(8M_W^2)$ one sees that $v = \sqrt{1/(\sqrt{2}G_F)}$.

The potential of the Higgs field is given by

$$V(\Phi^\dagger\Phi) = \mu^2(\Phi^\dagger\Phi) + \lambda(\Phi^\dagger\Phi)^2\,, \;\; \mu^2 < 0\,. \tag{12}$$

e) Y^u and Y^d are the so-called Yukawa couplings. Since a naive fermion mass term of the form $m(\bar\Psi_L\Psi_R + \bar\Psi_R\Psi_L)$ is not gauge invariant under $SU(2)_L$, the gauge-invariant Yukawa interaction was introduced to generate fermion masses. This is discussed in detail below.

1.3 The Yukawa interaction

After spontaneous symmetry breaking, the Yukawa term, in the case of only one fermion generation, contains fermion masses:

$$\begin{aligned}\mathcal{L}_{Yukawa} &= -\left(\bar\Psi_L\Phi^c Y_u u_R + \bar u_R \Phi^{c\dagger} Y_u^* \Psi_L\right) - \left(\bar\Psi_L \Phi Y_d d_R + \bar d_R \Phi^\dagger Y_d^* \Psi_L\right)\\ &\to -\frac{vY_u}{\sqrt2}\left(\bar u_L u_R + \bar u_R u_L\right) - \frac{vY_d}{\sqrt2}\left(\bar d_L d_R + \bar d_R d_L\right) + \ldots\,.\end{aligned} \tag{13}$$

In the last line we assumed that the Yukawa couplings are real, which leads to a simple mass term for the up- and down quarks with $m_{u,d} = vY_{u,d}/\sqrt2$. The possibility of having complex values of the Yukawa coupling is studied below.

For three generations of quarks the situation gets still a little more involved. The Yukawa interaction reads now

$$\begin{aligned}&\mathcal{L}_{Yukawa} =\\ -\;&\left(\bar Q_{1,L}, \bar Q_{2,L}, \bar Q_{3,L}\right)\Phi^c \hat Y_u \begin{pmatrix} u_R\\ c_R\\ t_R\end{pmatrix} - \left(\bar u_R, \bar c_R, \bar t_R\right)\Phi^{c\dagger}\hat Y_u^\dagger \begin{pmatrix} Q_{1,L}\\ Q_{2,L}\\ Q_{3,L}\end{pmatrix}\\ -\;&\left(\bar Q_{1,L}, \bar Q_{2,L}, \bar Q_{3,L}\right)\Phi \hat Y_d \begin{pmatrix} d_R\\ s_R\\ b_R\end{pmatrix} - \left(\bar d_R, \bar s_R, \bar b_R\right)\Phi^{\dagger}\hat Y_d^\dagger \begin{pmatrix} Q_{1,L}\\ Q_{2,L}\\ Q_{3,L}\end{pmatrix}\,,\end{aligned} \tag{14}$$

with the three $SU(2)_L$ doublets

$$Q_{1,L} = \begin{pmatrix} u_L\\ d_L\end{pmatrix}, \quad Q_{2,L} = \begin{pmatrix} c_L\\ s_L\end{pmatrix}, \quad Q_{3,L} = \begin{pmatrix} t_L\\ b_L\end{pmatrix}. \tag{15}$$

Note, that now in general the Yukawa coupling matrices $\hat Y_{u,d}$ do not have to be diagonal! After spontaneous symmetry breaking one gets the following structure of the fermion mass terms:

$$-\bar\Psi_L^u \hat M_1 \Psi_R^u - \bar\Psi_R^u \hat M_1^\dagger \Psi_L^u - \bar\Psi_L^d \hat M_2 \Psi_R^d - \bar\Psi_R^d \hat M_2^\dagger \Psi_L^d\,, \tag{16}$$

with

$$\Psi^u = \begin{pmatrix} u \\ c \\ t \end{pmatrix} , \qquad \Psi^d = \begin{pmatrix} d \\ s \\ b \end{pmatrix} , \tag{17}$$

$$\hat{M}_1 = \frac{v}{\sqrt{2}} \hat{Y}_u , \qquad \hat{M}_2 = \frac{v}{\sqrt{2}} \hat{Y}_d . \tag{18}$$

Again, in general the mass matrices $\hat{M}_1$ and $\hat{M}_2$ do not have to be diagonal, but they can be diagonalised with unitary transformations

$$\Psi^u \quad \rightarrow \quad U_1 \Psi^u \;\text{ with }\; U_1^\dagger U_1 = 1 , \tag{19}$$

$$\Psi^d \quad \rightarrow \quad U_2 \Psi^d \;\text{ with }\; U_2^\dagger U_2 = 1 . \tag{20}$$

The transformed mass matrices read

$$U_1^\dagger \hat{M}_1 U_1 = \frac{v}{\sqrt{2}} U_1^\dagger \hat{Y}_u U_1 = \begin{pmatrix} m_u & & \\ & m_c & \\ & & m_t \end{pmatrix} , \tag{21}$$

$$U_2^\dagger \hat{M}_2 U_2 = \frac{v}{\sqrt{2}} U_2^\dagger \hat{Y}_d U_2 = \begin{pmatrix} m_d & & \\ & m_s & \\ & & m_b \end{pmatrix} . \tag{22}$$

The corresponding **mass eigenstates** describe the **physical eigenstates**, whereas the original fields that couple to the weak gauge bosons define **weak eigenstates**. In principle the mass matrices could also be diagonal from the beginning. We will start, however, with the most general case and let the experimental data show what is actually realised in nature.

1.4 The CKM matrix

The transformation between weak and mass eigenstates does not affect the electromagnetic interaction and also not the neutral weak currents. In this cases up-like quarks couple among each other and similarly for the down-like quarks, so that only the combinations $U_1^\dagger U_1$ and $U_2^\dagger U_2$ arise in the interaction terms. By definition these combinations give the unit matrix. Thus all neutral interactions are diagonal, in other words **there are no flavour changing neutral currents (FCNC) in the Standard Model at tree-level.** The originally diagonal charged weak interaction can, however, become non-diagonal after performing the unitary transformations above, namely

$$\begin{aligned} & (\bar{u}, \bar{c}, \bar{t}) \, \gamma_\mu \, (1 - \gamma_5) \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \\ \rightarrow \quad & (\bar{u}, \bar{c}, \bar{t}) \, \gamma_\mu \, (1 - \gamma_5) \, U_1^\dagger U_2 \begin{pmatrix} d \\ s \\ b \end{pmatrix} \end{aligned}$$

$$\rightarrow \quad (\bar{u}, \bar{c}, \bar{t})\, \gamma_\mu \,(1-\gamma_5) \begin{pmatrix} .. & .. & .. \\ .. & .. & .. \\ .. & .. & .. \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} . \tag{23}$$

This defines the famous **Cabibbo-Kobayashi-Maskawa-Matrix** or **CKM-Matrix**

$$V_{CKM} := U_1^\dagger U_2 \,. \tag{24}$$

From a theory point of view it is not excluded that $U_1^\dagger U_2$ is diagonal (e.g. U_1 and U_2 are unit matrices or $U_1 = U_2$). In the end, it is the comparison with the experimental data that must indicate, as it has done, if the CKM-matrix is non-diagonal and thus transitions between different families are allowed. Historically this matrix was invented in two steps:

- 1963: 2×2 quark mixing matrix by Cabibbo [12]
- 1973: 3×3 quark mixing matrix by Kobayashi and Maskawa [13], who received the Nobel Prize in 2008.

Let us look a little closer at the properties of this matrix. By construction the CKM-Matrix is a unitary matrix, it connects the weak eigenstates q' with the mass eigenstates q. Note that instead of transforming both the up-type and down-type quark fields one can also solely transform the down-type fields, as

$$\begin{pmatrix} d \\ s \\ b \end{pmatrix} = V_{CKM} \begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} . \tag{25}$$

One can show, that a generic unitary $N \times N$-matrix has $N(N-1)/2$ real parameters and $(N-1)(N-2)/2$ phases, if unphysical phases are discarded. Specifically

N = 2	1 real parameter	0 phases
N = 3	3 real parameters	1 phase
N = 4	6 real parameters	3 phases

As will be discussed below, a complex coupling, e.g. a complex CKM-matrix element, leads to a phenomenon called **CP-violation**, which is strongly connected to the existence of matter in the Universe. Kobayashi and Maskawa found in 1973 that at least three families of quarks (i.e. six quarks) would be needed to implement CP-violation in the Standard Model. At that time only three quarks were known, the charm-quark was found in 1974 [14, 15].

The CKM-matrix allows for non-diagonal couplings in the charged currents, i.e. the u-quark does not only couple to the d-quark via a charged W boson, but it also couples to the s-quark and the b-quark, see Eq. (23). The entries of the CKM-matrix give the respective coupling strengths

$$V_{CKM} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} , \tag{26}$$

so that e.g. the coupling of an u- and d-quark is given by

$$\frac{g_2}{2\sqrt{2}}\gamma_\mu(1-\gamma_5)V_{ud}\,. \tag{27}$$

For a unitary 3×3 matrix with 3 real angles and 1 complex phase, different parameterisations are possible. The so-called **standard parameterisation** reads

$$V_{CKM} = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta_{13}} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta_{13}} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta_{13}} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta_{13}} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta_{13}} & c_{23}c_{13} \end{pmatrix}, \tag{28}$$

with

$$s_{ij} := \sin(\theta_{ij}) \quad \text{and} \quad c_{ij} := \cos(\theta_{ij})\,. \tag{29}$$

The three angles are θ_{12}, θ_{23} and θ_{13}, the complex phase describing CP-violation is δ_{13}. This parameterisation is exact and it is typically used for numerical calculations. There is also a very ostensive parameterisation, the so-called **Wolfenstein parameterisation** [16]. This parameterisation follows from the experimentally found hierarchy $V_{ud} \approx 1 \approx V_{cs}$ and $V_{us} \approx 0.22498 =: \lambda$ and is based on performing a Taylor expansion in λ. It is expressed in terms of 4 parameters λ, A, ρ and η, where the latter leads to complex contributions, i.e.

$$V_{CKM} = \begin{pmatrix} 1-\frac{\lambda^2}{2} & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\frac{\lambda^2}{2} & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}(\lambda^4)\,. \tag{30}$$

In this form the hierarchies can be read off very nicely. Transitions within a family are strongly favoured, transitions between the first and second family are suppressed by one power of λ, transitions between the second and third family are suppressed by two powers of λ and transitions between the first and the third family by at least three powers. The most recent numerical values for the Wolfenstein parameter read (status August 2024 from the CKMfitter page [17])

$$\lambda = 0.22498^{+0.00023}_{-0.00021}\,, \qquad A = 0.8215^{+0.0047}_{-0.0082}\,, \tag{31}$$

$$\bar{\rho} = \rho\left(1-\frac{\lambda^2}{2}\right) = 0.1562^{+0.0112}_{-0.0040}\,, \qquad \bar{\eta} = \eta\left(1-\frac{\lambda^2}{2}\right) = 0.3551^{+0.0051}_{-0.0057}\,. \tag{32}$$

Remarks:

– The non-vanishing value of η describes CP-violation within the Standard Model, which is by now unambiguously established.

– Numerically one gets $|V_{ub}| = 0.003730 = \lambda^{3.7482}$, so V_{ub} is closer to λ^4 than to λ^3, as it was historically assumed by Wolfenstein.

Numerically the moduli of the CKM matrix elements reads (status August 2024 from CKMfitter [17])

$$V_{CKM} = \begin{pmatrix} 0.974358^{+0.000049}_{-0.000054} & 0.22498^{+0.00023}_{-0.00022} & 0.003730^{+0.000044}_{-0.000048} \\ 0.22484^{+0.00023}_{-0.00021} & 0.973509^{+0.000054}_{-0.000059} & 0.04160^{+0.00020}_{-0.00058} \\ 0.008573^{+0.000046}_{-0.000158} & 0.04088^{+0.00020}_{-0.00066} & 0.9991248^{+0.0000268}_{-0.0000074} \end{pmatrix} . \tag{33}$$

Remarks:

- From these experimental numbers we can see clearly, that the CKM-matrix is non-diagonal. So the initial ansatz of non-diagonal Yukawa interactions was necessary!
- We can also clearly see the hierarchy of the CKM-matrix. Transitions within a family are strongly favoured, whereas transitions between different families are disfavoured. Note, however, that in the lepton sector there is a very different hierarchy.
- The above numbers have very small uncertainties. This relies crucially on the assumption of having a unitary 3×3 CKM matrix. Giving up this assumption, e.g. in models with four fermion generations, would lead to considerably larger uncertainties [18, 19], albeit the simplest versions of such models have already been experimentally excluded [20, 21].

1.5 Baryogenesis

In this section we discuss the important phenomenon of CP violation, namely the violation of the discrete symmetries of parity and charge conjugation, and how this is related to the origin of matter in the Universe, hence triggering a huge interest in both the theoretical and experimental communities. For more details, see the cosmology lecture of Prof. Shaposhnikov.

The observed asymmetry between matter and antimatter in the Universe can be parameterised by the baryon to photon ratio η_B, which was measured by the PLANCK satellite [22] to be

$$\eta_B = \frac{n_B - n_{\bar{B}}}{n_\gamma} \approx (6.05 \pm 0.07) \cdot 10^{-10} \, , \tag{34}$$

n_B is the number of baryons in the Universe, $n_{\bar{B}}$ the number of anti-baryons and n_γ denotes the number of photons. The tiny[2] matter excess is responsible for the whole visible Universe! In the very early Universe the relative excess of matter over antimatter was much smaller, compared to now, namely

$$\eta_B(t \approx 0) = \frac{10000000001 - 10000000000}{n_\gamma} \, , \tag{35}$$

$$\eta_B(today) = \frac{1-0}{n_\gamma} \, . \tag{36}$$

Sakharov has shown in 1967 [23] that one can create a baryon asymmetry dynamically (**Baryogenesis**), if the laws of nature have certain properties. These basics properties are:

[2] The numerical value is obtained by investigating primordial nucleosynthesis and the cosmic microwave background, see e.g. the PLANCK homepage.

a) **C and CP-violation:** C is the charge conjugation, it changes the sign of the charges of the elementary particles; P is the usual parity, a reflection of the three-dimensional space axes. The violation of parity in the weak interaction was established in 1956 [9] and in 1964 a tiny CP violation effect was found in the neutral K-system — in an observable denoted by ϵ — by Christenson, Cronin, Fitch, Turlay [24] (NP 1980). As will be discussed below, by now CP violation has by been observed in many processes involving b-hadrons, yielding also large effects of the order of 50% and recently also in the decays of D^0 mesons, where, however the effect appears to be tiny.
CP violation can be implemented in the SM via complex Yukawa-couplings, as with the CKM matrix, but also via complex parameters in the Higgs potential, as for example in the case with 2 Higgs doublet models, see e.g. [25].

b) **Baryon number violation:** The necessity to violate the baryon number is obvious and in the Standard Model such effects are implemented via so-called sphaleron processes — non-perturbative tunneling effects that arise at finite temperatures [26, 27].

c) **Phase out of thermal equilibrium:** This could be e.g. a first order phase transitions during electro-weak baryogenesis.

A remark for students:

- Sakharov's paper was sent to the journal on 23.9.1966 and published on 1.1.1967; it was cited for the first time in 1976 by Okun and Zeldovich; in summer 2024 it had almost 5000 citations ⇒ be patient with your papers!

All the three ingredients introduced above have to be part of the fundamental theory, not only in principal, but also to a sufficient extent.

a) In the Standard Model C and CP violation are implemented. A measure of the amount of CP violation is provided by the Jarlskog invariant J [28], which reads, in the Standard Model

$$J = (m_t^2 - m_c^2)(m_t^2 - m_u^2)(m_c^2 - m_u^2)(m_b^2 - m_s^2)(m_b^2 - m_d^2)(m_s^2 - m_d^2) \cdot A \,. \tag{37}$$

Here, m_q denotes the mass of the quark q and A the area of the unitarity triangle, which will be discussed below. A is large if the CKM-matrix elements have large imaginary parts, i.e. in the presence of large CP violating contributions. Normalising J to the scale of the electroweak phase transition leads to the very small number, see e.g. Ref. [29]:

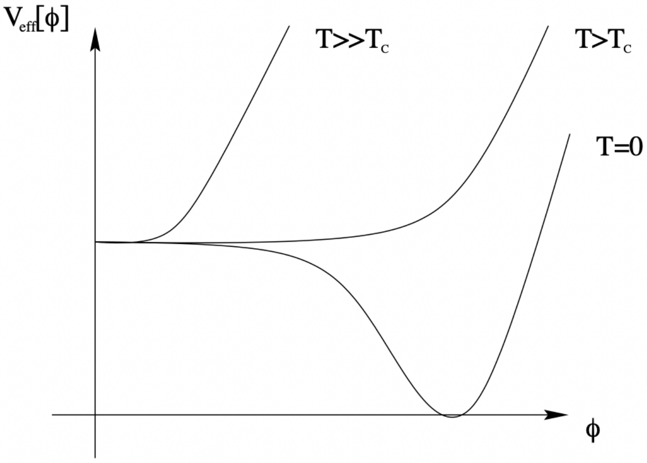
$$\frac{J}{(100\,\mathrm{GeV})^{12}} \approx 10^{-20} \ll 6 \times 10^{-10} \approx \eta_B \,. \tag{38}$$

Hence, it seems that the amount of CP violation in the Standard Model is by far not sufficient to explain the observed baryon asymmetry. Note, however, that if one could let baryogenesis take place at a lower energy scale, e.g. 10 GeV, then the above ratio would be enhanced by a factor 10^{12} and the amount of CP violation in the Standard Model could be sufficient. Such a possibility is investigated e.g. in Ref. [30].

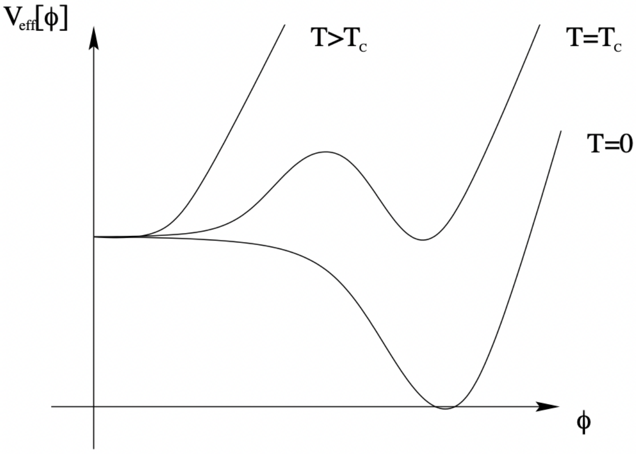
b) In the Standard Model the baryon number (B) and the lepton number (L) are conserved to leading order in perturbation theory. Considering also non-perturbative effects (there are no corresponding

Feynman diagrams!), in particular thermal effects, it is possible to create the needed amount of violation of B. These effects are called *sphalerons* (greek: weak, dangerous) [26, 27]. At temperatures T$<$ 100 GeV, this phenomenon is exponentially suppressed, while it grows very rapidly above 100 GeV. Sphalerons have not yet been detected experimentally.

c) Finally one needs to be out of thermal equilibrium at 100 GeV. During a second order phase transition the parameters change in a continuous way and there is no departure from thermal equilibrium:



In order to leave thermal equilibrium a first order transition is needed:



To answer the question about the nature of the electroweak phase transition one has to calculate the effective Higgs potential (classical potential plus quantum effects) as a function of the Higgs mass at finite temperature. In particular, one finds, that for masses m_H $<$ 72 GeV a first order transition is possible, while the transition is continuous for higher masses, see e.g. Refs. [31–34]. Thus, the experimental value of the Higgs mass of 125 GeV clearly points towards a continuous phase transition within the Standard Model. We note again, that 2HDM models could also fulfill this Sakharov criteria and lead to a first order phase transition, see e.g. Refs. [35, 36].

For lecture notes on baryogenesis see e.g. Ref. [37].

1.6 Unitarity Triangle

Next we discuss in more detail the determination of the CKM-matrix and in particular the so-called unitarity triangle. By construction we have a unitary CKM matrix, i.e.

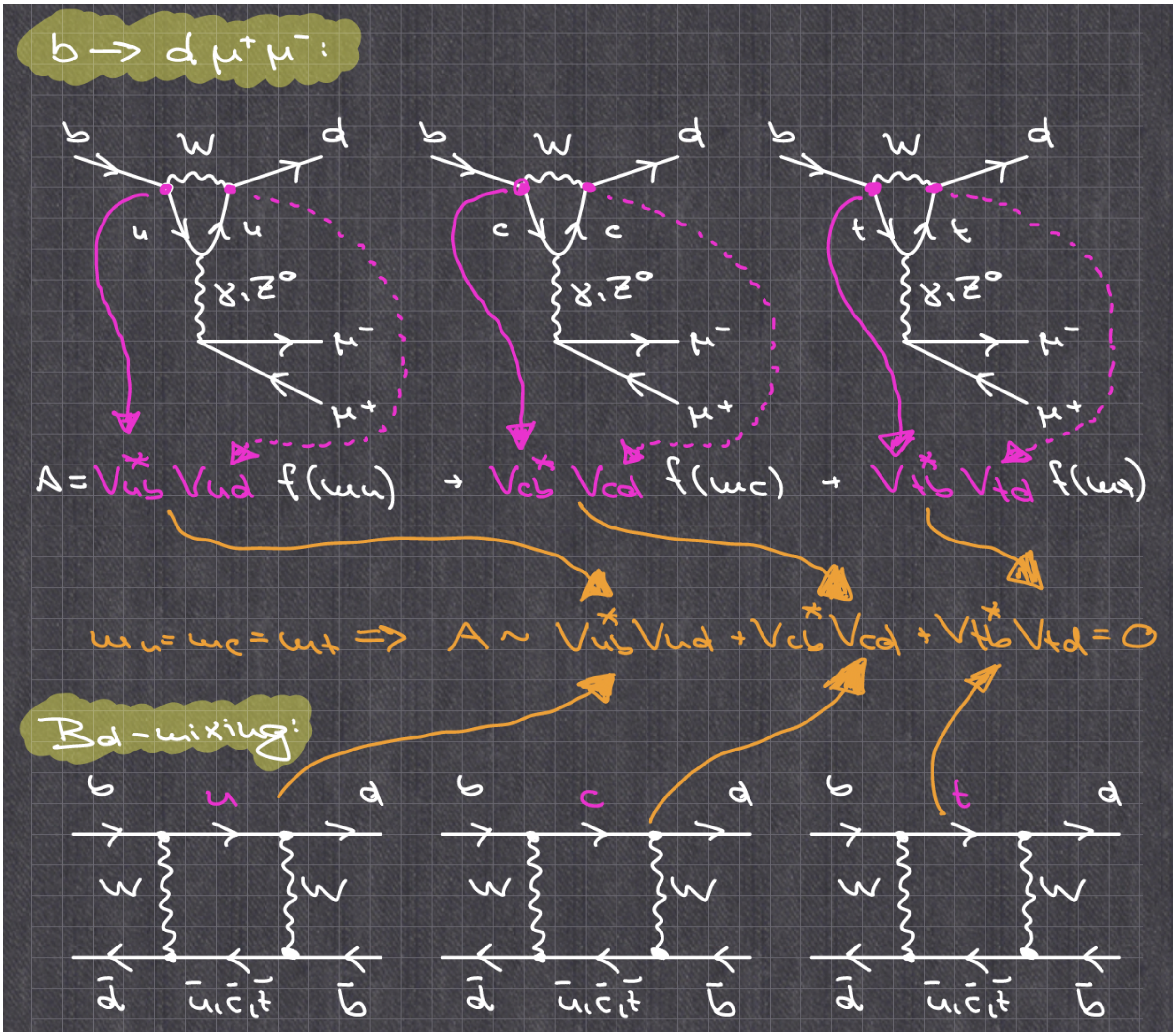
$$1 \;=\; V_{CKM}^{\dagger} V_{CKM} = \sum_{U=u,c,t} V_{Ud_1}^{*} V_{Ud_2} = \begin{pmatrix} 1 & 0_{K^0} & 0_{B_d} \\ 0 & 1 & 0_{B_s} \\ 0 & 0 & 1 \end{pmatrix} , \tag{39}$$

$$1 \;=\; V_{CKM} V_{CKM}^{\dagger} = \sum_{D=d,s,b} V_{u_1 D} V_{u_2 D}^{*} = \begin{pmatrix} 1 & 0_{D^0} & 0_{T^0} \\ 0 & 1 & 0_{T_c^0} \\ 0 & 0 & 1 \end{pmatrix} . \tag{40}$$

We will explain below, what the subscripts on the zero entries of the unit matrix mean.
Each of the unitarity conditions in Eq. (39) and Eq. (40) leads to nine equations, specifically three combinations of CKM matrix elements with sum equal to one and six combinations with sum equal to zero. In particular, the condition

$$V_{ud} V_{ub}^{*} + V_{cd} V_{cb}^{*} + V_{td} V_{tb}^{*} = 0_{B_d} , \tag{41}$$

arises in the penguin diagrams triggering the $b \to d\mu^+\mu^-$ decay or in B_d mixing (i.e. the transition of a B_d meson into its anti-particle, the $\bar{B}_d$ meson, via a so-called box diagrams), see the figure below, when it is assumed that the masses of the internal quarks (up, charm and top) are equal.


b → d μ⁺μ⁻ :
b
W
d
u u
c c
t t
γ, Z⁰
μ⁻
μ⁺
A = V*ub Vud f(mu) + V*cb Vcd f(mc) + V*tb Vtd f(mt)
mu = mc = mt ⇒ A ~ V*ub Vud + V*cb Vcd + V*tb Vtd = 0
Bd-mixing:
b
u
c
t
d
W
W
d̄
ū, c̄, t̄
b̄

Since the masses of the up-, charm- and top-quark are in reality different from each other, the above observation means that finite contributions to the $b \to d\mu^+\mu^-$ penguin decay and B_d mixing are only arising due to the mass difference of the internal quarks. This is the essence of the so-called **GIM-mechanism** (Glashow-Iliopoulos-Maiani) [38].

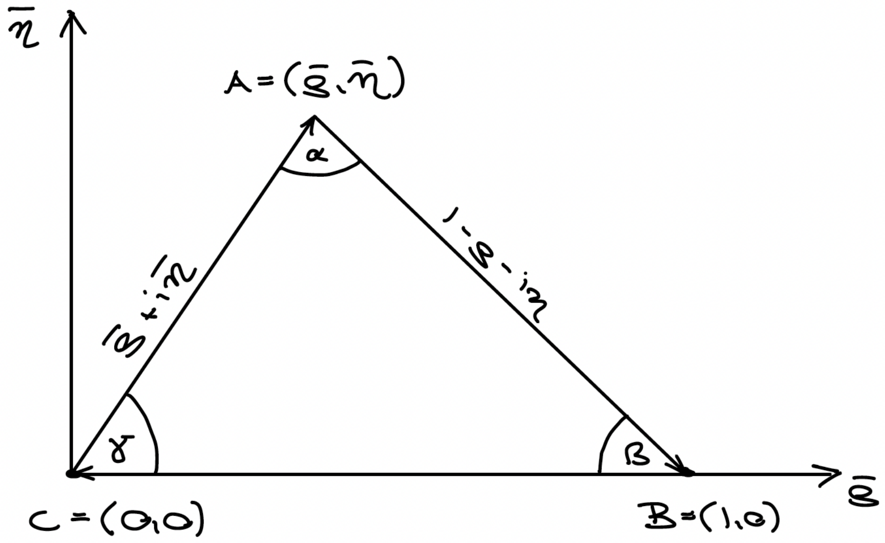
The remaining zero-sums in Eq. (39) and Eq. (40) correspond to $s \to d$ penguins (or K^0 mixing), $b \to s$ penguins (or B_s mixing), $c \to u$ penguins (or D^0 mixing), $t \to u$ penguins and $t \to c$ penguins.

Using the Wolfenstein parameterisation we get for the B_d zero sum

$$A\lambda^3 \left[(\bar{\rho} + i\bar{\eta}) - 1 + (1 - (\rho + i\eta))\right] = 0 + \mathcal{O}(\lambda^4)\,. \tag{42}$$

As the values for A and λ are already quite precisely known, Eq. (42) can be used to determine ρ and η. The above sum of three complex numbers can be represented graphically as a triangle in the complex (ρ, η) plane, the so-called *unitarity triangle*[3], see the figure below

[3] In principle different unitarity triangles, apart from the B_d one, defined in Eq. 42, can be constructed, but they turn out to be very "flat", i.e. they have one very small angle, while in the B_d unitarity triangle all three lengths are of similar size.



The precise determination of the unitarity triangle given by Eq. 42 is of particular interest since a non-vanishing value of η is a measure of the size of CP-violation in the Standard Model.

In order to constrain the form of the unitarity triangle, the following strategy is used (for an early review see e.g. [39]): to compare the experimental value of some flavour observables with the corresponding theoretical expressions, where ρ and η are left as free parameters, plot the constraints on these two parameters in the complex (ρ,η) plane (in this case the values of λ and A are assumed to be known and fixed) e.g.:

- The amplitude describing the $b \to u$ decay is proportional to V_{ub}. Therefore the branching fraction for the semileptonic decay of a $\bar{B}$-meson (containing a b-quark) into a meson containing a u-quark, i.e. X_u, is proportional to $|V_{ub}|^2$:

$$\begin{aligned} B^{\text{exp.}}(B \to X_u e \overline{\nu}) &= \tilde{a}^{\text{th.}} \cdot |V_{ub}|^2 = a^{\text{th.}} \cdot \left(\rho^2+\eta^2\right) , \\ \Rightarrow \rho^2+\eta^2 &= \frac{B^{\text{exp.}}(B \to X_u e \overline{\nu})}{a^{\text{th.}}} , \end{aligned} \tag{43}$$

  where $a^{\text{th.}}$ contains the result of the theoretical calculation. Comparing experiment and theory for this observable and leaving ρ and η as free parameters, leads to a constraint in the (ρ,η)-plane in the form of a circle around $(0,0)$ with the radius $B^{\text{exp.}}(B \to X_u e \overline{\nu})/a_{\text{th.}}$.

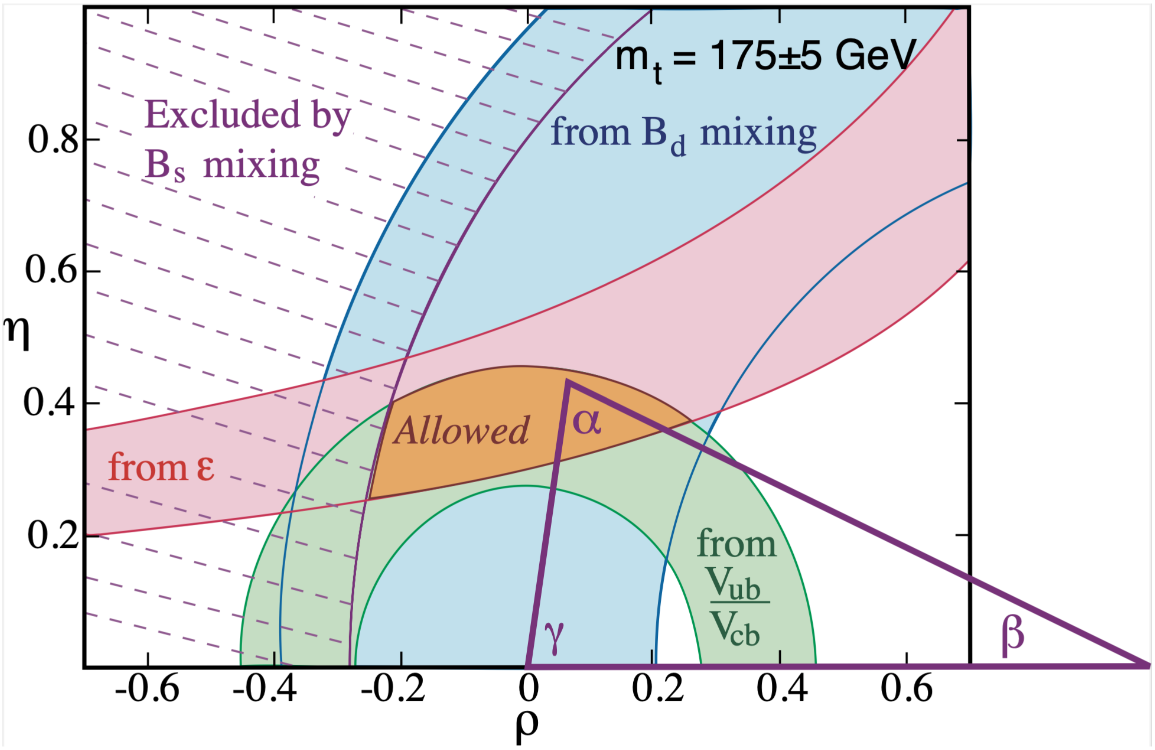
- Investigating the dynamics of neutral B_d-mesons, one finds that the physical eigenstates (i.e. the states that are propagating with a definite mass) are a mixture of the flavour eigenstates (defined by the quark content of the neutral mesons), as it will be discussed in more detail below. As a result of this mixing, the two physical eigenstates have different masses and their difference is denoted by ΔM_{B_d}. Theoretically one has to determine the above shown box-diagrams and one finds to a very good approximation that only the CKM structure stemming from the internal top-quark is contributing. Therefore the following relation holds:

$$\Delta M_{B_d}^{\text{exp.}} \propto |V_{td}|^2 \propto (\rho-1)^2+\eta^2 , \tag{44}$$

which leads to a circle in the (ρ, η)-plane around $(1, 0)$, when comparing experiment and theory.

- Comparing theory and experiment for CP-violation in the neutral K-meosn system, denoted by ϵ, gives an hyperbola in the (ρ, η)-plane.

The overlap of these three regions constrains the values of ρ and η, as schematically shown in the figure below, with the bound from the semileptonic B decay denoted in green, the constraint from B-mixing in blue and the hyperbolic constraint from ϵ in pink.



This figure is just meant to visualise the method in principle, below we show a plot with the latest experimental numbers.

Remarks:

- The angles of the unitarity triangle can be expressed as

$$\alpha = \arg\left(-\frac{V_{td}V_{tb}^*}{V_{ud}V_{ub}^*}\right), \qquad \beta = \arg\left(-\frac{V_{cd}V_{cb}^*}{V_{td}V_{tb}^*}\right), \qquad \gamma = \arg\left(-\frac{V_{ud}V_{ub}^*}{V_{cd}V_{cb}^*}\right). \tag{45}$$

The length of the sides of the unitarity triangle can be related to the angles α, β, γ, as

$$\overline{AC} = R_b = \frac{\sin\beta}{\sin(\beta+\gamma)}, \qquad \overline{AB} = R_t = \frac{\sin\gamma}{\sin(\beta+\gamma)}. \tag{46}$$

To a good approximation we can also write

$$V_{ub} = |V_{ub}|e^{-i\gamma}, \qquad V_{td} = |V_{td}|e^{-i\beta}. \tag{47}$$

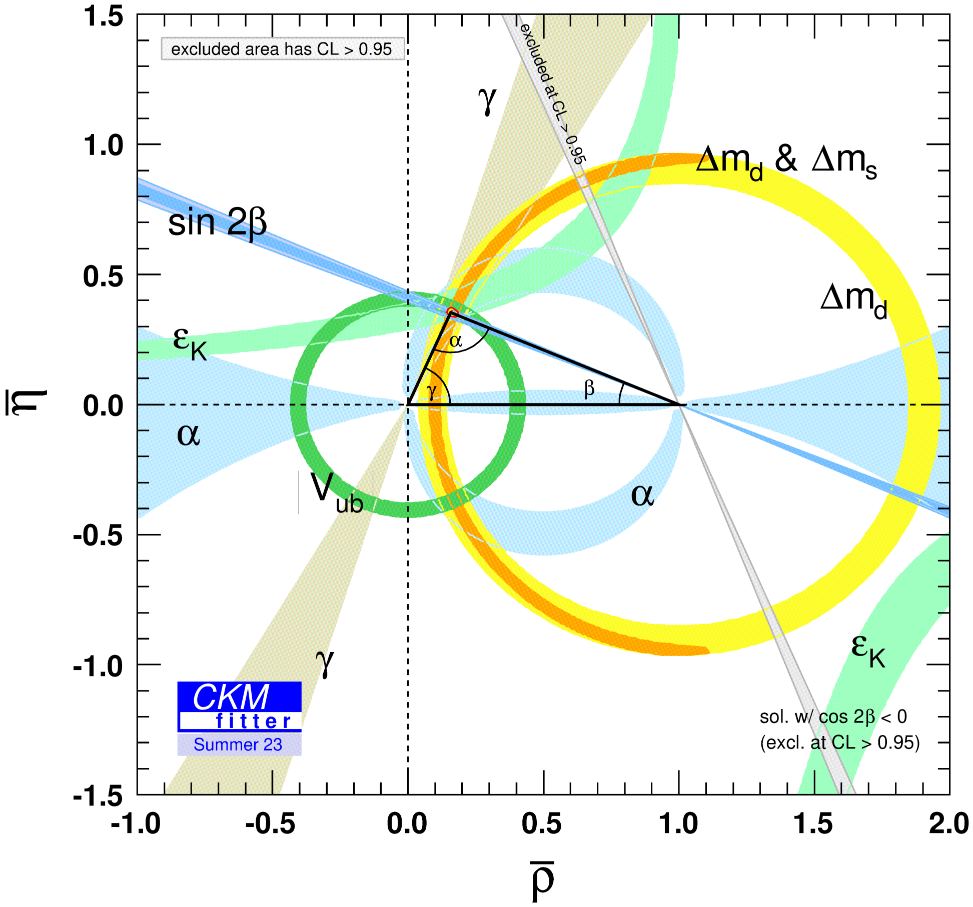
Bigi and Sanda have shown that the angle β can be extracted directly, with almost no theoretical uncertainty — thus it is called *gold-plated mode* — from the following CP-asymmetry in exclusive

B-meson decays [40].

$$a_{CP} \quad := \quad \frac{\Gamma(B \to J/\Psi + K_S) - \Gamma(\bar{B} \to J/\Psi + K_S)}{\Gamma(B \to J/\Psi + K_S) + \Gamma(\bar{B} \to J/\Psi + K_S)} \propto \sin 2\beta \tag{48}$$

The fact that the size of $\sin 2\beta$ was expected to be of order one was a strong reason for building the **B-factories** in SLAC (with the detector BaBar) and at KEK (with the detector Belle(II)) to measure for the first time CP violation outside the Kaon sector, which was achieved in 1999 [41, 42]. Currently a huge experimental effort is put into the direct determination of the CKM angle γ. Assuming no BSM effects in hadronic tree-level decays, this extraction can be done with essentially no (of the order of 10^{-6}) uncertainties [43].[4] The determination of the CKM angle α suffers from more pronounced theory uncertainties.

– The above programme was performed in the last years with great success and it confirmed the CKM picture, see below the most recent contraints for the unitarity triangle from the CKMfitter group [17]. Similar results have been obtained by the UTfit collaboration [46]. As a result of these efforts Kobayashi and Maskawa were awarded the Nobel Prize in 2008.

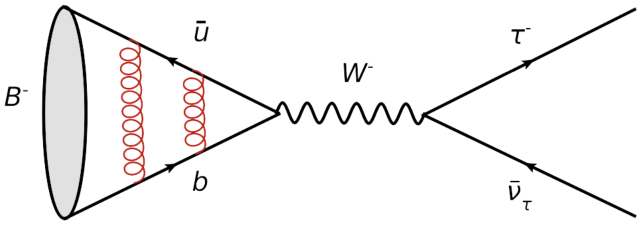


[4] Giving up the assumption of having no BSM effects in tree-level non-leptonic decays can lead to very large shifts in the determination of the CKM angle γ [44, 45].

1.7 Classification of hadronic decays

1.7.1 Leptonic decays

Leptonic decays have only leptons in the final state, e.g. the tree-level decay $B^- \to \tau^- \, \bar{\nu}_\tau$.



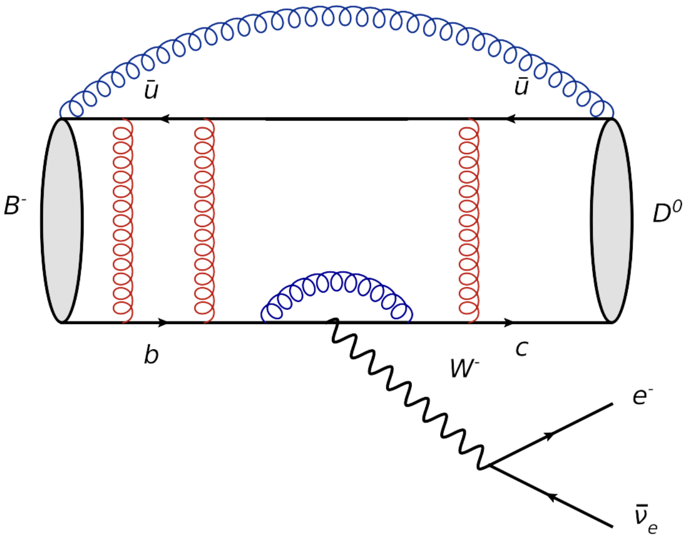
Such decays have the simplest hadronic structure. Gluons bind the quarks of the initial state into a hadron. All non-perturbative effects are described by one parameter: the **decay constant**, f_{B^-}, which is defined for generic B mesons as

$$\langle 0|\bar{q}\gamma^\mu\gamma_5 b|B_q(p)\rangle \quad = \quad i f_{B_q} p^\mu \,, \tag{49}$$

where b and u are the spinors of the bottom and up quark and p^μ is the B_q-meson four-momentum. Decay constants can nowadays be precisely determined by lattice QCD simulations. Note that leptonic decays can also proceed via loop-level contributions in the SM, an example is the decay $B_s \to \mu^+\mu^-$.

1.7.2 Semi-leptonic decays

Semi-leptonic decays have leptons and hadrons in the final state, e.g. the tree-level decay $B^- \to D^0 \, e^- \, \bar{\nu}_e$.



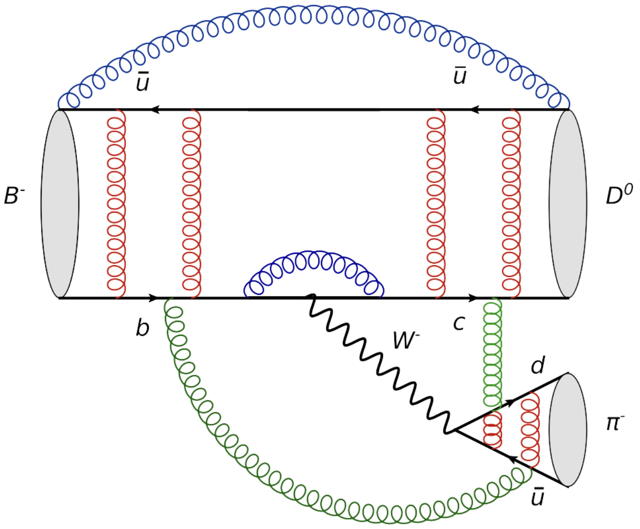
Now the hadronic structure is more complicated, as the non-perturbative QCD-effects are not only responsible for the binding of the hadrons in the initial and final states, but also for the strong interaction between the initial and the final state. The non-perturbative dynamics, in this case, is described in terms of two functions, the **form factors** $f_+^{B^- \to D^0}(q^2)$ and $f_0^{B^- \to D^0}(q^2)$, that depend on the momentum transfer q^2. They are defined as

$$\begin{aligned}\langle D^0(p_D)|\bar{c}\gamma^\mu b|B^-(p_B)\rangle = \quad & f_+^{B^- \to D^0}(q^2)\left(p_B^\mu + p_D^\mu - \frac{m_B^2 - m_D^2}{q^2}q^\mu\right) \\ & + f_0^{B^- \to D^0}(q^2)\frac{m_B^2 - m_D^2}{q^2}q^\mu \,. \end{aligned} \tag{50}$$

Form factors can be determined by sum rules or Lattice QCD calculations. There are again semi-leptonic decays that can only proceed via loop-contributions in the SM, e.g. $B_d \to K^{0*}\mu^+\mu^-$, which belong to the class of decays, where currently deviations between experiment and theory are observed.

1.7.3 Non-leptonic decays

Non-leptonic decays have only hadrons in the final state, e.g. the tree-level decay $B^- \to D^0\,\pi^-$.



These are the most complicated decays and they can only be described theoretically by making additional assumptions that then allow for a factorisation, e.g.

$$\begin{aligned}&\langle D^0\pi^-|\bar{c}\gamma_\mu(1-\gamma_5)b \cdot \bar{d}\gamma^\mu(1-\gamma_5)u|B^-\rangle \\ &\quad \approx \langle D^0|\bar{c}\gamma_\mu(1-\gamma_5)b|B^-\rangle \quad \cdot \quad \langle \pi^-|\bar{d}\gamma^\mu(1-\gamma_5)u|0\rangle \\ &\quad \propto f^{B^- \to D^0}(q^2 = m_\pi^2) \quad \cdot \quad f_\pi \,. \end{aligned} \tag{51}$$

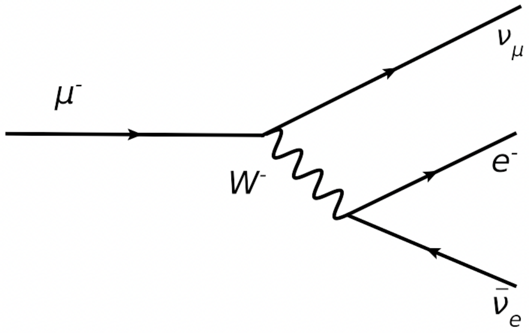
Theoretical investigations, of when the factorisation assumption is justified and when not, are a hot topic of current research, see e.g. [47].

2 Lecture 2: Theoretical framework

In comparison to the first and third lecture, this lecture will be considerably more technical, in particular the main part, which is devoted to the effective Hamiltonian.

2.1 The Muon Decay

The muon decay $\mu^- \to \nu_\mu\; e^-\; \bar{\nu}_e$ represents the most simple type of weak decay, because there are no QCD effects involved.[5] This process is given by the following Feynman diagram.



The total decay rate of the muon reads (see e.g. [48] for an early reference)

$$\Gamma_{\mu\to\nu_\mu+e+\bar{\nu}_e} = \frac{G_F^2 m_\mu^5}{192\pi^3} f\left(\frac{m_e}{m_\mu}\right) = \frac{G_F^2 m_\mu^5}{192\pi^3} c_{3,\mu}\,. \tag{52}$$

$G_F = g_2^2/(4\sqrt{2}M_W^2)$ denotes the Fermi constant and f the phase space factor for one massive particle in the final state. It is given by

$$f(x) \quad = \quad 1 - 8x^2 + 8x^6 - x^8 - 24x^4 \ln(x) = c_{3,\mu}\,. \tag{53}$$

The coefficient $c_{3,\mu}$ is introduced here to be consistent with the later notation. The result in Eq. (52) is already very instructive, since from that we obtain that the measurable lifetime of the muon which reads

$$\tau = \frac{1}{\Gamma} = \frac{192\pi^3}{G_F^2 m_\mu^5 f\left(\frac{m_e}{m_\mu}\right)}\,. \tag{54}$$

Thus the lifetime of a weakly decaying particle is proportional to the inverse of the fifth power of its mass. Using the measured values [49] for $G_F = 1.1663788(6)\cdot 10^{-5}$ GeV^{-2}, $m_e = 0.51099895000(15)$ MeV and $m_\mu = 0.1056583755(23)$ GeV, we can predict[6] the lifetime of the muon to be

$$\tau_\mu^{\text{th.}} = 2.18776\cdot 10^{-6}\,\text{s}\,, \tag{55}$$

[5] This statements holds to a high accuracy. QCD effects arise for the first time at the two loop order.

[6] This is of course not really correct, because the measured muon lifetime was used to determine the Fermi constant, but for pedagogical reasons we assume that the Fermi constant is already known.

which is in excellent agreement with the measured value [49] of

$$\tau_\mu^{\text{exp.}} = 2.1969811(22) \cdot 10^{-6}\,\text{s}\;. \tag{56}$$

The remaining tiny difference (the prediction is about 0.4% smaller than the experimental value) is due to higher order electro-weak corrections. These are crucial for a highly precise determination of the Fermi constant. The dominant contribution is given by the 1-loop QED corrections, calculated already in the 1950s [50, 51]:

$$c_{3,\mu} = f\left(\frac{m_e}{m_\mu}\right)\left[1 + \frac{\alpha}{4\pi}2\left(\frac{25}{4} - \pi^2\right)\right]\;. \tag{57}$$

Taking this effect into account ($\alpha = 1/137.035999074(44)$ [49]) we then obtain

$$\tau_\mu^{\text{th.}} = 2.19699 \cdot 10^{-6}\,\text{s}\;, \tag{58}$$

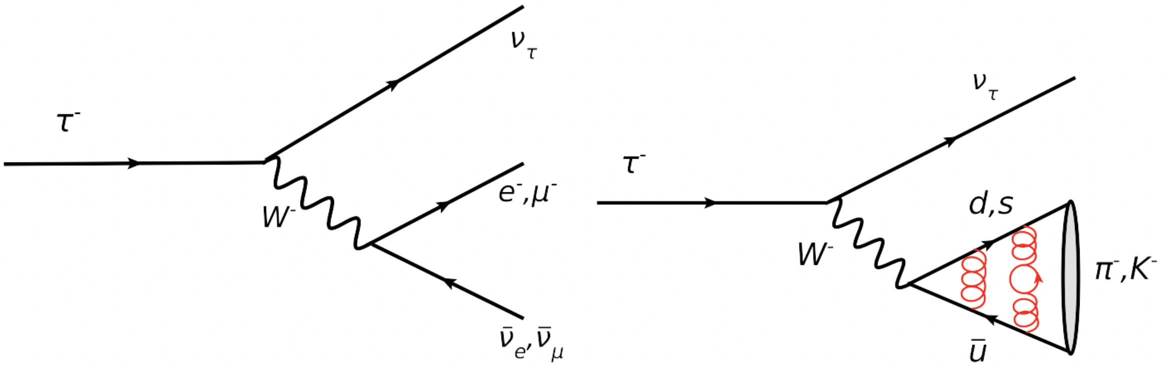
which almost exactly reproduces the measured value given in Eq. (56). The complete 2-loop QED contributions have been determined in [52], while a review of loop-corrections to the muon decay is given in [53].

Note that the phase space factor gives an almost negligible suppression in the specific case of the muon decay - namely $f(m_e/m_\mu) = 0.999813 = 1 - 0.000187051$. However, phase space effects will turn out to be quite sizable, e.g. for the decay of a b-quark into a c-quark.

2.2 The tau decay

Moving to the tau lepton, both decays into leptons as well as into quarks and leptons are possible, specifically

$$\tau \quad\to\quad \nu_\tau + \begin{cases} e^- + \bar{\nu}_e \\ \mu^- + \bar{\nu}_\mu \\ d + \bar{u} \\ s + \bar{u} \end{cases} \qquad .$$



Note that heavier quarks, like the charm- or bottom-quark cannot be created, since in this case the lightest meson possible, i.e. $D^0 = c\bar{u}$ with $M_{D^0} \approx 1.86\,\text{GeV}$ is heavier than the tau lepton

$(m_\tau = 1.77682(16)$ GeV). Thus the total decay rate of the tau lepton reads

$$\begin{aligned} \Gamma_\tau &= \frac{G_F^2 m_\tau^5}{192\pi^3}\left[f\left(\frac{m_e}{m_\tau}\right) + f\left(\frac{m_\mu}{m_\tau}\right) + N_c\,|V_{ud}|^2\, g\left(\frac{m_u}{m_\tau},\frac{m_d}{m_\tau}\right) + N_c\,|V_{us}|^2\, g\left(\frac{m_u}{m_\tau},\frac{m_s}{m_\tau}\right)\right] \\ &=: \frac{G_F^2 m_\tau^5}{192\pi^3} c_{3,\tau}\,. \end{aligned} \tag{59}$$

Here, $N_c = 3$ is a colour factor and we have introduced the new phase space function g, which accounts for two massive particles in the final state. If we neglect the phase-space effects due to the lepton and quark masses ($f(m_e/m_\tau) = 1 - 7\cdot 10^{-7}; f(m_\mu/m_\tau) = 1 - 0.027; ...$) and if we use $V_{ud}^2 + V_{us}^2 \approx 1$, we obtain $c_{3,\tau} = 5$ and thus the simple approximate relation

$$\frac{\tau_\tau}{\tau_\mu} = \left(\frac{m_\mu}{m_\tau}\right)^5 \frac{1}{5}\,. \tag{60}$$

Using the experimental values for τ_μ, m_μ and m_τ we can predict

$$\tau_\tau^{\text{th.}} = 3.26707\cdot 10^{-13}\ \text{s}\,, \tag{61}$$

which is quite close to the experimental value of

$$\tau_\tau^{\text{exp.}} = 2.906(1)\cdot 10^{-13}\ \text{s}\,. \tag{62}$$

Note that the theory prediction is about 12% larger than the measured value. This is mostly due to the effect of sizable QCD corrections, which must be taken into account since now there are quarks in the final state - contrary to the case of the muon decay. QCD contributions currently have been calculated up to five loop accuracy [54], a review of higher order corrections can be found in [55].

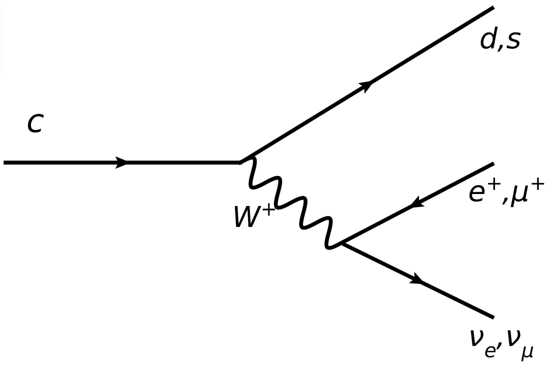
Because of the pronounced and clean dependence on the strong coupling, tau decays can also be used for precision determinations of α_s, see, e.g., the review [56].

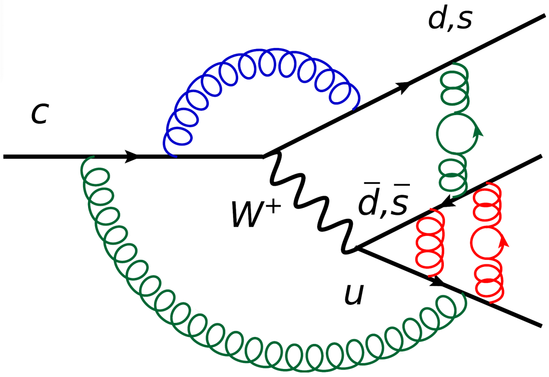
This example already shows, that a proper treatment of QCD effects is mandatory for precise lifetime studies. In the case of meson decays this will be even more important.

2.3 Charm-quark decay

Before trying to investigate the complicated structure of meson decays, let us have a look at the decay of free c- and b-quarks. Within the framework of the **Heavy Quark Expansion (HQE)** one can show that the free quark decay is the leading term in a systematic expansion in the inverse of the heavy (decaying) quark mass. For a both pedagogical and technical introduction into the HQE, see [57], for a review with many historic details, see [58] and for a review covering the state of the art of HQE predictions and comparisons to experiment, see [59].

A charm quark can decay weakly into a strange- or a down-quark and a W^+-boson, which can then further decay either into leptons (semi-leptonic decay) or into quarks (non-leptonic decay).





Calculating the inclusive total decay rate of a charm-quark we obtain

$$\Gamma_c = \frac{G_F^2 m_c^5}{192\pi^3} |V_{cs}|^2 c_{3,c} \, , \tag{63}$$

with

$$\begin{aligned} c_{3,c} = \quad & g\left(\frac{m_s}{m_c}, \frac{m_e}{m_c}\right) + g\left(\frac{m_s}{m_c}, \frac{m_\mu}{m_c}\right) + N_c |V_{ud}|^2 h\left(\frac{m_s}{m_c}, \frac{m_u}{m_c}, \frac{m_d}{m_c}\right) \\ & \qquad + N_c |V_{us}|^2 h\left(\frac{m_s}{m_c}, \frac{m_u}{m_c}, \frac{m_s}{m_c}\right) \\ & + \left|\frac{V_{cd}}{V_{cs}}\right|^2 \left\{ g\left(\frac{m_d}{m_c}, \frac{m_e}{m_c}\right) + g\left(\frac{m_d}{m_c}, \frac{m_\mu}{m_c}\right) + N_c |V_{ud}|^2 h\left(\frac{m_d}{m_c}, \frac{m_u}{m_c}, \frac{m_d}{m_c}\right) \right. \\ & \qquad \left. + N_c |V_{us}|^2 h\left(\frac{m_d}{m_c}, \frac{m_u}{m_c}, \frac{m_s}{m_c}\right) \right\} \, . \end{aligned} \tag{64}$$

Here h denotes a new phase space function, for the case of three massive particles in the final state. If we set all phase space factors to one ($f(m_s/m_c) = f(0.0935/1.471) = 1 - 0.03, \ldots$ with $m_s = 93.5(2.5)$ MeV [49]) and use $|V_{ud}|^2 + |V_{us}|^2 \approx 1 \approx |V_{cd}|^2 + |V_{cs}|^2$, then we get $c_{3,c} = 5$, similar to the τ decay. In that case we can predict a charm lifetime of

$$\tau_c \quad = \quad \left\{ \begin{array}{l} 0.84 \text{ ps} \\ 1.70 \text{ ps} \end{array} \right. \quad \text{for } m_c = \left\{ \begin{array}{lll} 1.471 & \text{GeV} & \text{(Pole-scheme)} \\ 1.277(26) & \text{GeV} & (\overline{MS} - \text{scheme}) \end{array} \right. \, . \tag{65}$$

These predictions lie roughly in the ball-bark of the experimental numbers for D-meson lifetimes, but at this stage some comments are mandatory:

- Predictions of the lifetimes of free quarks have a huge parametric dependence on the definition of the quark mass ($\propto m_q^5$). This is the reason, why for a long time only lifetime ratios (the dominant m_q^5 dependence as well as CKM factors and some sub-leading non-perturbative corrections cancel) were determined theoretically. In our case the value obtained in the $\overline{MS}$ – scheme for the charm quark mass is about a factor of 2 larger than the one obtained in the pole-scheme. At LO-QCD the definition of the quark mass is completely arbitrary, which leads to huge uncertainties. Performing the computation at NLO-QCD a consistent treatment of the quark masses has to be defined within the calculation, leading to a considerably weaker dependence of the final result on the quark mass definition.

- Taking only the decay of the free c-quark into account, one obtains the same lifetimes for all charm-mesons, which is clearly a very bad approximation, taking the large spread of lifetimes of different D-mesons into account.

$$\frac{\tau(D^+)}{\tau(D^0)} = 2.54(2)\,, \qquad \frac{\tau(D_s^+)}{\tau(D^0)} = 1.20(1)\,. \tag{66}$$

  Within the framework of the HQE one will find that in the case of charmed mesons a very sizable contribution to the total decay rate stems from spectator effects where also the valence quark of the D-meson is involved in the decay.
- Perturbative QCD corrections to the free charm quark decay turn out to be very important, because $\alpha_s(m_c) \approx 0.35$ is quite large.

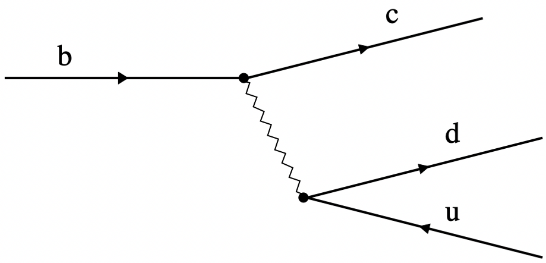
In the framework of the HQE the spectator effects will turn out to be suppressed by $1/m_c^3$ and since m_c is not very large, the suppression is also not expected to be very pronounced. This will change in the case of B-mesons. Because of the larger value of the b-quark mass, one expects a better description of the meson decay in terms of the simple b-quark decay, which is confirmed by experiment

$$\frac{\tau(B^+)}{\tau(B_d^0)} = 1.076(4)\,, \qquad \frac{\tau(B_s^0)}{\tau(B_d^0)} = 1.002(4)\,. \tag{67}$$

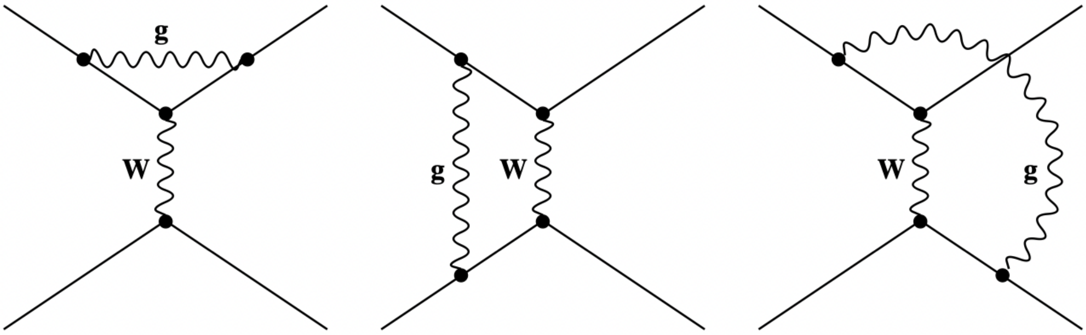
In the next lecture we will have a closer look into how one can determine perturbative QCD contributions to the quark and meson decay in a consistent and effective way - this will lead to the concept of the **effective Hamiltonian**. A nice pedagogical introduction to this topic is given in [60].

2.4 The effective Hamiltonian

Most weak decays proceed via the exchange of heavy W-bosons, e.g. the decay $b \to c + W^- \to c + \bar{u} + d$ is described by the following Feynman diagram,

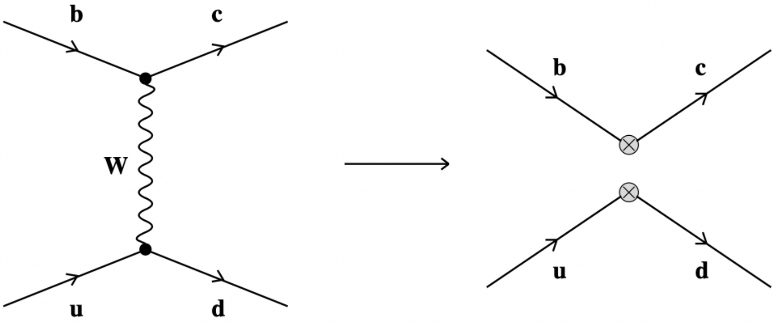


Neglecting the masses of the final state quarks, two very different scales arise in this decay: the mass of the W-boson (≈ 80 GeV) and the mass of the b-quark (≈ 4.5 GeV). Since the strong coupling is sizeable at the b quark scale, $\alpha_s(m_b) \approx 0.2$, perturbative QCD corrections are expected to be important. Specifically, 1-loop diagrams will give α_s corrections, 2-loop diagrams $\alpha_s^2 \approx 0.04$ corrections, and so on. Calculating the 1-loop QCD corrections one finds, however, that besides terms of order α_s, also **large logarithmic terms** of the form $\alpha_s \ln\left(m_b^2/M_W^2\right)$ appear.



As a result we do not obtain a Taylor expansion in α_s but an expansion in $\alpha_s \ln\left(m_b^2/M_W^2\right) \approx 6\alpha_s$ which clearly spoils the perturbative approach.

The solution to this problem lies in the introduction of the **effective Hamiltonian**, described in some detail below. The basic idea is to derive an effective theory valid at scales of the order of m_b, in which the heavy W-boson ($m_W \gg m_b$), that triggers the weak decay, is **integrated out** by performing an operator product expansion (OPE), see, e.g., [60] for a nice introduction, as well as [61]. Schematically in the resulting effective theory, the propagator of the W-boson is contracted to a point and new effective four-fermion operators are generated, see figure below.



The Feynman rules for the diagram on the l.h.s. give the following expression, which can be Taylor-expanded in k^2/M_W^2.

$$\bar{c}\frac{ig_2V_{cb}^*}{2\sqrt{2}}\gamma^\mu(1-\gamma_5)b\frac{1}{k^2-M_W^2}\bar{d}\frac{ig_2V_{ud}}{2\sqrt{2}}\gamma_\mu(1-\gamma_5)u\,. \tag{68}$$

Note, that the momentum transfer in the virtual W-boson is of the order of the b-quark mass, $k \approx m_b$. The leading term of the Taylor-expansion of this expression in $k^2/M_W^2 \approx 3.6\cdot 10^{-3}$, gives the effective Hamiltonian in zeroth order in the strong coupling.[7]

$$\begin{aligned}\Rightarrow \mathcal{H}_{eff}(x) &= \left(\frac{g_2}{2\sqrt{2}}\right)^2\frac{1}{M_W^2}V_{cb}^*V_{ud}\bar{c}\gamma^\mu(1-\gamma_5)b\cdot\bar{d}\gamma_\mu(1-\gamma_5)u \\ &=: \frac{G_F}{\sqrt{2}}V_{CKM}C_2Q_2\,. \end{aligned} \tag{69}$$

[7] Expanding the W-propagator in coordinate space, leads to a delta function with the difference of the coordinates of the two arising quark currents as an argument. Thus this approximation yields an effective local operator.

We have introduced the following notation: the four quark operator is denoted by Q_2

$$\begin{aligned} Q_2 &= (\bar{c}_\alpha \gamma_\mu (1-\gamma_5) b_\alpha) \cdot \left(\bar{d}_\beta \gamma^\mu (1-\gamma_5) u_\beta\right) , \\ &=: (\bar{c}_\alpha b_\alpha)_{V-A} \cdot \left(\bar{d}_\beta u_\beta\right)_{V-A} , \end{aligned} \tag{70}$$

where α and β denote colour indices, this operator is a colour-singlet. as well as the Fermi constant G_F with

$$G_F = \frac{g_2^2}{4\sqrt{2} M_W^2} = 1.1663787(6) \cdot 10^{-5} \frac{1}{\mathrm{GeV}^2} , \tag{71}$$

thus leading to the **Fermi-theory** of the weak interaction. We stress that so far the **Wilson coefficient** $C_2 = 1$, however, including also QCD corrections the above description is generalised as follows:

- The value of $C_2 = 1 + \mathcal{O}(\alpha_s)$ deviates from one and it depends on the renormalisation scale μ. For b and c-decays its value is slightly larger than one.
- A second colour-rearranged operator Q_1 arises.

$$Q_1(x) =: (\bar{c}_\alpha b_\beta)_{V-A} \cdot \left(\bar{d}_\beta u_\alpha\right)_{V-A} . \tag{72}$$

The value of $C_1 = \mathcal{O}(\alpha_s)$ is negative for b and c-decays and of the order of -20% for b-decays.

- The generic effective Hamiltonian for tree-level decays reads thus

$$\mathcal{H}_{eff}(x) = \frac{G_F}{\sqrt{2}} V_{CKM} \left[C_1(\mu) Q_1(x) + C_2(\mu) Q_2(x)\right] . \tag{73}$$

- As mentioned above in the Standard Model large logarithms arise in the perturbative expansion and in the end one will not have an expansion in the strong coupling, $\alpha_s(m_b) \approx 0.2$, but an expansion in $\ln(m_b/M_W)^2 \alpha_s \approx 1.2$. So the convergence of the expansion is not ensured. In particular, the general structure of the perturbative expansion reads

	LL	NLL	NNLL	NNNLL
Tree	1	–	–	–
1-loop	$\alpha_s \ln$	α_s	–	–
2-Loop	$\alpha_s^2 \ln^2$	$\alpha_s^2 \ln$	α_s^2	–
3-loop	$\alpha_s^3 \ln^3$	$\alpha_s^3 \ln^2$	$\alpha_s^3 \ln$	α_s^3
	...	...	...	...

(74)

Computations within the full Standard Model correspond to calculating row by row, whereas using the effective Hamiltonian corresponds to calculating column by column. Importantly, the latter framework also allows to **sum up the large logarithms to all orders**.[8] An example for such a summation is given by the solution of the renormalisation group equations for the strong coupling, which is discussed e.g. in the lecture notes in [62].

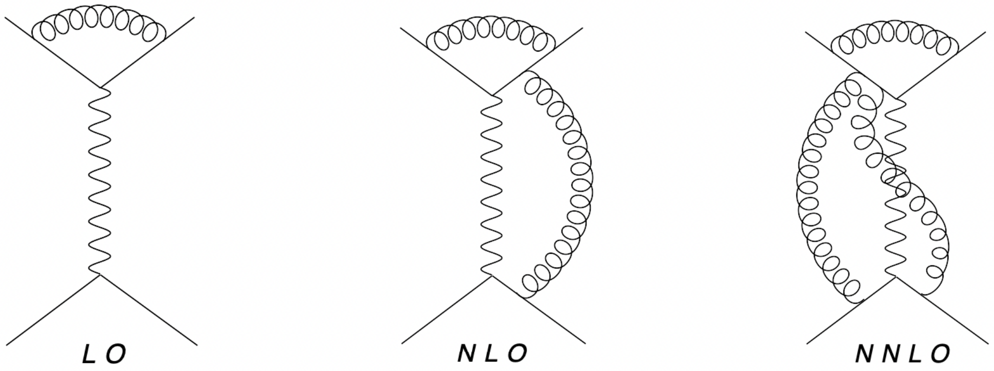
- In meson decays besides perturbatively calculable short-distance (SD) QCD effects (e.g. those at the scale M_W), also long-distance (LD) effects arise (at the scale Λ_{QCD}), the latter being of

[8] LL = leading logarithmic approximation, NLL = next-to-leading logarithmic approximation,...

non-perturbative origin. The construction of an OPE leads to a well-defined **separation of scales**. Namely the high energy physics is described by the Wilson coefficients, which can be calculated in perturbation theory, whereas the hadronic effects are parametrised by the matrix element of the operators Q_1, Q_2. In this case non-perturbative methods like lattice QCD or QCD sum rules must be employed. The renormalisation scale μ acts as the factorisation scale.

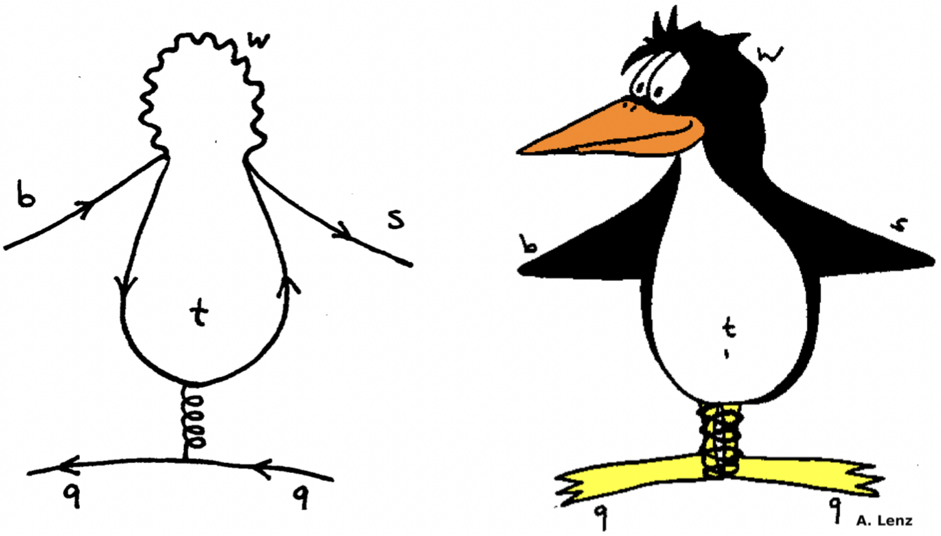
- Calculations within the framework of the effective Hamiltonian are technically simpler, because fewer propagators appear in the expressions.

Historic remark: The effective Hamiltonian in Eq. (73) was already obtained in 1974 in LL-QCD [63], a nice review of the NLL-results is given in [61]. Currently, also NNLL results are available [64]. To obtain the expression of the Wilson coefficients at LO-QCD 1-loop diagrams have to be calculated, at NLO-QCD 2-loop corrections and for NNLO-QCD 3-loop corrections, i.e.



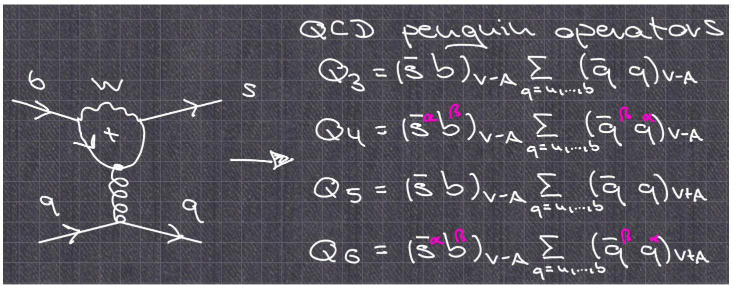
2.5 Penguins and friends

So far we have only considered tree-level decays, however there are also loop-induced decays, like those described by the penguins diagram[9].



[9] For the origin of the name penguin diagram have a look at page 5 of [65].

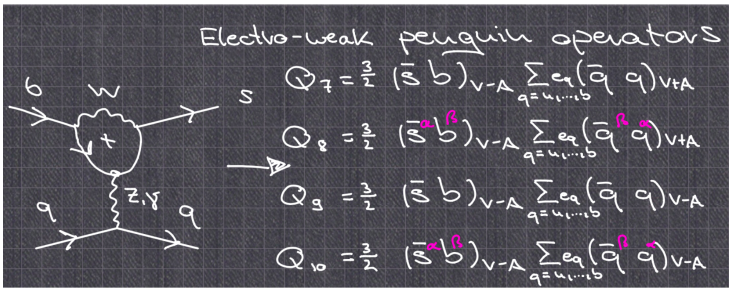
Here again, the particles heavier than the bottom quark — now in addition to the W, also the Z and the top-quark can appear — are integrated out, yielding new, effective operators, such as the **QCD-penguin operators** $Q_{3,\ldots,6}$, explicitly given in the figure below.



QCD penguin operators

$$Q_3 = (\bar{s} b)_{V-A} \sum_{q=u,\ldots,b} (\bar{q} q)_{V-A}$$
$$Q_4 = (\bar{s}^{\alpha} b^{\beta})_{V-A} \sum_{q=u,\ldots,b} (\bar{q}^{\beta} q^{\alpha})_{V-A}$$
$$Q_5 = (\bar{s} b)_{V-A} \sum_{q=u,\ldots,b} (\bar{q} q)_{V+A}$$
$$Q_6 = (\bar{s}^{\alpha} b^{\beta})_{V-A} \sum_{q=u,\ldots,b} (\bar{q}^{\beta} q^{\alpha})_{V+A}$$

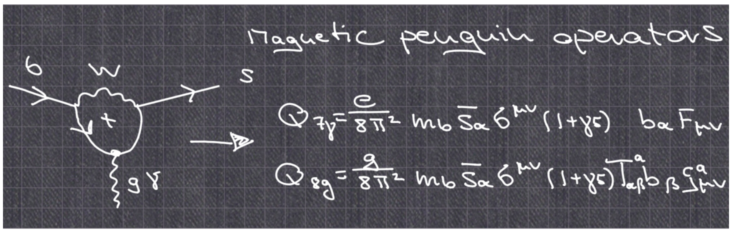
The QCD-penguin operators contribute e.g. to the decays $b \to c\bar{c}s$, where they give sizable corrections and to $b \to u\bar{u}s$, where they can even dominate. **Electro-weak penguin operators** (in which the gluon is replaced by a photon or a Z-boson) are denoted by $Q_{7,\ldots,10}$, see below.



Electro-weak penguin operators

$$Q_7 = \frac{3}{2} (\bar{s} b)_{V-A} \sum_{q=u,\ldots,b} e_q (\bar{q} q)_{V+A}$$
$$Q_8 = \frac{3}{2} (\bar{s}^{\alpha} b^{\beta})_{V-A} \sum_{q=u,\ldots,b} e_q (\bar{q}^{\beta} q^{\alpha})_{V+A}$$
$$Q_9 = \frac{3}{2} (\bar{s} b)_{V-A} \sum_{q=u,\ldots,b} e_q (\bar{q} q)_{V-A}$$
$$Q_{10} = \frac{3}{2} (\bar{s}^{\alpha} b^{\beta})_{V-A} \sum_{q=u,\ldots,b} e_q (\bar{q}^{\beta} q^{\alpha})_{V-A}$$

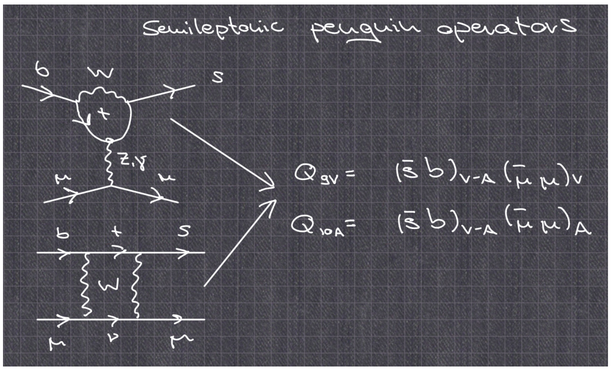
Note that the latter operators typically give tiny corrections, however, they give the dominant contribution to the direct CP violation in the Kaon system, namely to ϵ'/ϵ. **Magnetic penguin operators**, see below,



Magnetic penguin operators

$$Q_{7\gamma} = \frac{e}{8\pi^2} m_b \bar{s}_\alpha \sigma^{\mu\nu} (1+\gamma_5) b_\alpha F_{\mu\nu}$$
$$Q_{8g} = \frac{g}{8\pi^2} m_b \bar{s}_\alpha \sigma^{\mu\nu} (1+\gamma_5) T^a_{\alpha\beta} b_\beta G^a_{\mu\nu}$$

correspond to penguin diagrams in which the photon ($Q_{7\gamma}$) or the gluon is on-shell (Q_{8g}), i.e. its four momentum squared is zero. $Q_{7\gamma}$ gives the leading contribution to the radiative decay $b \to s\gamma$. Finally we have **semi-leptonic penguin operators**, that are generated either by penguin or by box diagrams, see below.


Semileptonic penguin operators
b
W
s
t
Z,γ
μ
μ
b
t
s
W
μ
ν
μ
$Q_{9V} = (\bar{s}b)_{V-A}(\bar{\mu}\mu)_V$
$Q_{10A} = (\bar{s}b)_{V-A}(\bar{\mu}\mu)_A$

Q_{10A} gives the main contribution to the decay $B_s \to \mu\mu$, while in semi-leptonic decays like $B \to K^*\mu^+\mu^-$ also Q_{9V} contributes. Note that often the same notation is used for these operators as for the electro-weak penguins.

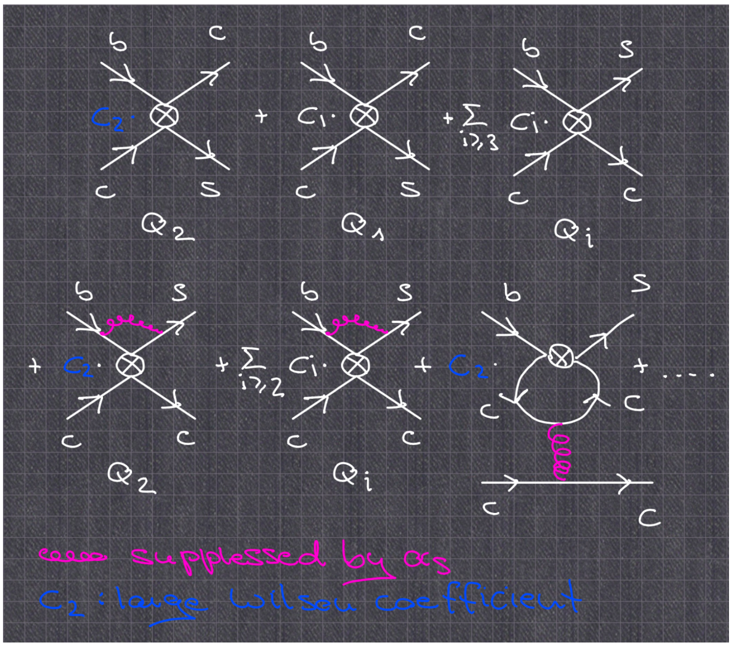
In the end we arrive at the complete expression of the **effective weak Hamiltonian**:

$$\mathcal{H}_{eff} = \frac{G_F}{\sqrt{2}} \left[\sum_{q=u,c} V_c^q \left\{ C_1(\mu) Q_1^q + C_2(\mu) Q_2^q \right\} - V_p \sum_{j \geq 3} C_j(\mu) Q_j \right] + \text{h.c.}\,. \tag{75}$$

The Vs denote different combinations of CKM elements, the operators $Q_3, ..$ denote all the different penguin operators discussed above. The values of the numerically leading tree-level Wilson coefficients C_1 and C_2 have been determined above, the QCD penguin Wilson coefficients are below 5%, with the exception of C_{8g}, the coefficient of the chromomagnetic operator and electro-weak penguins are even smaller.

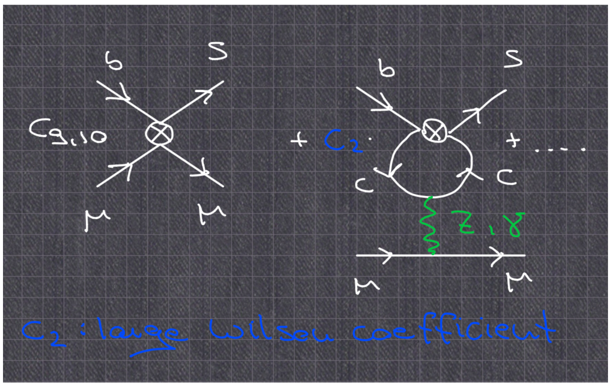
Having the effective Hamiltonian at hand, we can now calculate different processes — forgetting about the underlying weak structure of the SM — using the Wilson coefficients as basic couplings of our theory and the four quark operators as basic vertices.

For a calculation of the rate $\Gamma(b \to c\bar{c}s)$, we have to consider the following contributions:



The leading contribution is of course the tree-level insertion of the operator Q_2, the tree-level insertion of Q_1 and Q_{3-6} leads to smaller corrections. QCD effects are also sizable. The dominant one originates from QCD-corrections to the tree-level insertion of color-singlet operator Q_2. Note that at this order, also the penguin insertion of Q_2 is expected to be sizable.

Finally, to compute the decay rate $\Gamma(b \to s\mu^+\mu^-)$, where currently some anomalies are observed, we have to consider the following contributions:



Here the dominant one comes from the tree-level insertion of the operators $Q_{9,10}$. Due to the very large Wilson coefficient, we also expect, however, sizable corrections from the penguin insertion of the operator Q_2. These are the so-called charm-loop effects, often discussed in the context of the flavour anomalies, see e.g. [66].

3 Lecture 3: Mixing and CP violation

After having introduced the concept of the effective Hamiltonian, which is the starting point for calculating decay rates of exclusive or inclusive decays, we will switch now the topic and give a brief introduction into the concepts of mixing and CP violation. More detailed discussions of mixing and CP violation can be found e.g. in the reviews [67–69].

3.1 General introduction

In these lecture, we mostly discuss the case of B_d mesons; changes in formulae in order to describe B_s mesons or D^0 mesons follow straightforwardly. However, when important differences arise, we will discuss each system separately.

Neutral mesons like the B_d^0 and its anti particle $\bar{B}_d^0$ form a two-states system, which can be described by a Schrödinger-like equation[10], as

$$i\hbar \frac{\partial}{\partial t} \begin{pmatrix} B_d^0 \\ \bar{B}_d^0 \end{pmatrix} = \hat{H} \begin{pmatrix} B_d^0 \\ \bar{B}_d^0 \end{pmatrix} = \begin{pmatrix} M_{11}^d - \frac{i}{2}\Gamma_{11}^d & 0 \\ 0 & M_{22}^d - \frac{i}{2}\Gamma_{22}^d \end{pmatrix} \begin{pmatrix} B_d^0 \\ \bar{B}_d^0 \end{pmatrix} . \tag{76}$$

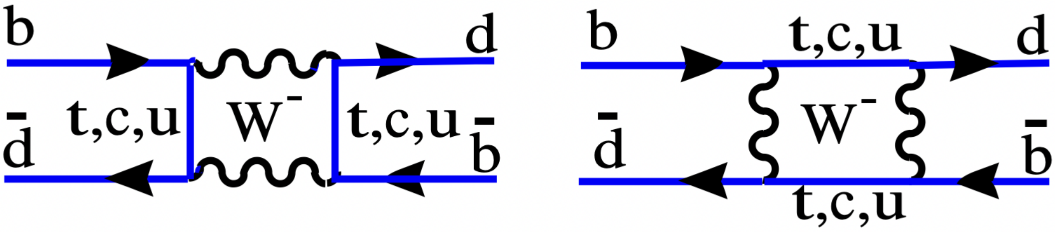
[10] Since in this case the Hamiltonian is not hermitian, we do not call it Schrödinger-equation, even if the mathematical structure of the differential equation is identical to the usual Schrödinger-equation.

This leads to the following time evolution of B mesons:

$$\Rightarrow B_i(t) \;=\; e^{\frac{1}{i\hbar}\left(M^d_{ii}-\frac{i}{2}\Gamma^d_{ii}\right)t} = e^{\frac{1}{i\,\hbar}M^d_{ii}t}e^{-\frac{1}{2\,\hbar}\Gamma^d_{ii}t}\,. \tag{77}$$

- $M^d_{11}(M^d_{22})$ is the mass of the $B^0_d(\bar{B}^0_d)$-meson.
- $\Gamma^d_{11}(\Gamma^d_{22})$ is the total decay rate of the $B^0_d(\bar{B}^0_d)$-meson.
- CPT invariance implies $M^d_{11} = M^d_{22}$ and $\Gamma^d_{11} = \Gamma^d_{22}$.

Due to the weak interaction, however, transitions of a B^0_d-meson into a $\bar{B}^0_d$ (and vice versa) are possible via the so-called **box diagrams**, as shown below.



The box diagrams generate off-diagonal entries in the Hamiltonian, e.g.

$$\hat{H} \;=\; \begin{pmatrix} M^d_{11}-\frac{i}{2}\Gamma^d_{11} & M^d_{12}-\frac{i}{2}\Gamma^d_{12} \\ M^d_{21}-\frac{i}{2}\Gamma^d_{21} & M^d_{22}-\frac{i}{2}\Gamma^d_{22} \end{pmatrix}. \tag{78}$$

Γ^d_{12} corresponds to intermediate on-shell states, like $(c\bar{c})$, in the box-diagrams, while M^d_{12} corresponds to virtual, i.e. off-shell, intermediate states. Therefore the top quark as well as other potential heavy new physics particles contribute only to M^d_{12}. Thus both the mass and decay rate matrices are non-diagonal, simply meaning that the flavour eigenstates (defined by the quark content) of the two-meson system are not mass eigenstates (defined by the mass of the physical meson).

CPT invariance implies again that $M^d_{11} = M^d_{22}$ and $\Gamma^d_{11} = \Gamma^d_{22}$, while hermiticity gives $M^d_{21} = \left(M^d_{12}\right)^*$ and $\Gamma^d_{21} = \left(\Gamma^d_{12}\right)^*$.

By diagonalising the 2×2 Hamiltonian matrix of Eq. (78), we obtain mass eigenstates, labelled respectively as "heavy" (H) and "light" (L), that is

$$\begin{aligned} B_{d,H} &= pB^0_d - q\bar{B}^0_d\,, \\ B_{d,L} &= pB^0_d + q\bar{B}^0_d\,, \end{aligned} \tag{79}$$

with $p = p(M^d_{12}, \Gamma^d_{12})$ and $q = q(M^d_{12}, \Gamma^d_{12})$. The new eigenstates $B_{d,H}$ and $B_{d,L}$ have now definite masses M^d_H, M^d_L and definite decay rates Γ^d_H and Γ^d_L, leading to a decay rate difference $\Delta\Gamma_d$ and a mass difference ΔM_d.

$$\begin{aligned} \Delta\Gamma_d &= \Gamma^d_L - \Gamma^d_H = \Delta\Gamma_d(M^d_{12}, \Gamma^d_{12})\,, \\ \Delta M_d &= M^d_H - M^d_L = \Delta M_d(M^d_{12}, \Gamma^d_{12})\,, \end{aligned} \tag{80}$$

which are observable in experiment. The following relations hold exactly

$$(\Delta M_d)^2 - \frac{1}{4}(\Delta\Gamma_d)^2 = 4\left|M_{12}^d\right|^2 - \left|\Gamma_{12}^d\right|^2 \,, \tag{81}$$

$$\Delta M_d \cdot \Delta\Gamma_d = -4\mathrm{Re}\left(M_{12}^d \Gamma_{12}^{d*}\right) \,, \tag{82}$$

$$\frac{q}{p} = -\frac{\Delta M_d + \frac{i}{2}\Delta\Gamma_d}{2M_{12}^d - i\Gamma_{12}^d} \,. \tag{83}$$

Solving for the mass and decay rate difference gives

$$\begin{aligned} 2\Delta M_d^2 &= \sqrt{\left(4\left|M_{12}^d\right|^2 - \left|\Gamma_{12}^d\right|^2\right)^2 + 16\left|M_{12}^d\right|^2\left|\Gamma_{12}^d\right|^2 \cos^2\phi_{12}^d} + 4\left|M_{12}^d\right|^2 - \left|\Gamma_{12}^d\right|^2 \,, \\ \frac{1}{2}\Delta\Gamma_d^2 &= \sqrt{\left(4\left|M_{12}^d\right|^2 - \left|\Gamma_{12}^d\right|^2\right)^2 + 16\left|M_{12}^d\right|^2\left|\Gamma_{12}^d\right|^2 \cos^2\phi_{12}^d} - 4\left|M_{12}^d\right|^2 + \left|\Gamma_{12}^d\right|^2 \,, \end{aligned} \tag{84}$$

with the mixing phase $\phi_{12}^d = \arg(-M_{12}^d/\Gamma_{12}^d)$.[11] Hence, in general, both M_{12}^d and Γ_{12}^d need to be known in order to determine either the decay rate difference or the mass difference.

One finds, however, that one can make the approximations $|\Gamma_{12}^d/M_{12}^d| \ll 1$ in the B_d-system (as well as in the B_s-system) and $|\phi_{12}^D| \ll 1$ in the D-system, leading to the simplified relations:

$$\Delta M_d = 2|M_{12}^d| \,, \quad \Delta\Gamma_d = 2|\Gamma_{12}^d|\cos(\phi_{12}^d) \qquad \text{for } B_d \text{ mixing} \,, \tag{85}$$

$$\Delta M_s = 2|M_{12}^s| \,, \quad \Delta\Gamma_s = 2|\Gamma_{12}^s|\cos(\phi_{12}^s) \qquad \text{for } B_s \text{ mixing} \,, \tag{86}$$

$$\Delta M_D = 2|M_{12}^D| \,, \quad \Delta\Gamma_D = 2|\Gamma_{12}^D| \qquad \text{for } D \text{ mixing} \,. \tag{87}$$

3.2 Time evolution

Now we can derive in a similar way, as in the well-known example of neutrino oscillations (see also the lectures of Prof. Barenboim), the time evolution of the B_d mesons. For the mass eigenstates, the time evolution is trivial, i.e.

$$|B_{d,H/L}(t)\rangle = e^{-\left(iM_{H/L}^d + \Gamma_{H/L}^d/2\right)t}|B_{d,H/L}(0)\rangle \,. \tag{88}$$

For the flavour eigenstates it reads

$$|B_d^0(t)\rangle = g_+(t)|B_d^0\rangle + \frac{q}{p}g_-(t)|\bar{B}_d^0\rangle \,, \tag{89}$$

$$|\bar{B}_d^0(t)\rangle = \frac{p}{q}g_-(t)|B_d^0\rangle + g_+(t)|\bar{B}_d^0\rangle \,, \tag{90}$$

with the coefficients

$$g_+(t) = e^{-i\cdot M_{B_d}t}e^{-\Gamma_{B_d}/2t}\left[\cosh\frac{\Delta\Gamma_d t}{4}\cos\frac{\Delta M_d t}{2} - i\sinh\frac{\Delta\Gamma_d t}{4}\sin\frac{\Delta M_d t}{2}\right] \,, \tag{91}$$

$$g_-(t) = e^{-iM_{B_d}\cdot t}e^{-\Gamma_{B_d}/2t}\left[-\sinh\frac{\Delta\Gamma_d t}{4}\cos\frac{\Delta M_d t}{2} + i\cosh\frac{\Delta\Gamma_d t}{4}\sin\frac{\Delta M_d t}{2}\right] \,. \tag{92}$$

[11] With $M_{12}^d = |M_{12}^d|e^{i\phi_{M^d}}$ and $\Gamma_{12}^d = |\Gamma_{12}^d|e^{i\phi_{\Gamma^d}}$ we get $\phi_{12}^d = \pi + \phi_{M^d} - \phi_{\Gamma^d}$.

Here we introduced the averaged masses $M_{B_d^0}$ and decay rates Γ:

$$M_{B_d} = \frac{M_H^d + M_L^d}{2} \, , \qquad \Gamma_{B_d} = \frac{\Gamma_H^d + \Gamma_L^d}{2} \, . \tag{93}$$

The coefficient functions $g_+(t)$ and $g_-(t)$ give the probability for mixing to occur, namely

$$\begin{aligned} \left| \langle B_d^0 | B_d^0(t) \rangle \right|^2 &= |g_+(t)|^2 = \left| \langle \bar{B}_d^0 | \bar{B}_d^0(t) \rangle \right|^2 \, , & (94) \\ \left| \langle \bar{B}_d^0 | B_d^0(t) \rangle \right|^2 &= \left| \frac{q}{p} \right|^2 |g_-(t)|^2 \, . & (95) \end{aligned}$$

The arguments of the trigonometric and hyperbolic functions in Eq. (91) and Eq. (92) can be rewritten as

$$\begin{aligned} \frac{\Delta M_d \cdot t}{2} &= \frac{1}{2} x(B_d^0) \frac{t}{\tau(B_d^0)} \quad \text{with } x(B_d^0) := \frac{\Delta M_d}{\Gamma_d} \, , & (96) \\ \frac{\Delta \Gamma_d \cdot t}{4} &= \frac{1}{2} y(B_d^0) \frac{t}{\tau(B_d^0)} \quad \text{with } y(B_d^0) := \frac{\Delta \Gamma_d}{2\Gamma_d} \, , & (97) \end{aligned}$$

where the lifetime $\tau(B_d)$ is related to the total decay rate Γ_d via $\tau(B_d) = 1/\Gamma_{B_d}$. The oscillation length of the trigonometric functions can be determined via

$$\begin{aligned} \frac{\Delta M_d \cdot t}{2} &= \pi \Rightarrow t = \frac{2\pi}{\Delta M_d} & (98) \\ \Rightarrow x &= vt' = \beta\gamma ct = \beta\gamma \frac{2\pi c}{\Delta M_d} \, . & (99) \end{aligned}$$

Next we can also write down the time dependent decay rate of a B_d meson, that was initially (at time $t = 0$) tagged as a B_d flavour eigenstate into an arbitrary final state f:

$$\begin{aligned} \Gamma \left[B_d(t) \to f \right] = N_f \left| \mathcal{A}_f \right|^2 \left(1 + |\lambda_f|^2 \right) e^{-\Gamma t} \quad & \left\{ \frac{\cosh \left(\frac{\Delta \Gamma_d}{2} t \right)}{2} + \frac{1 - |\lambda_f|^2}{1 + |\lambda_f|^2} \frac{\cos \left(\Delta M_d t \right)}{2} \right. \\ - \frac{2\Re(\lambda_f)}{1 + |\lambda_f|^2} \frac{\sinh \left(\frac{\Delta \Gamma_d}{2} t \right)}{2} \quad & \left. - \frac{2\Im(\lambda_f)}{1 + |\lambda_f|^2} \frac{\sin \left(\Delta M_d t \right)}{2} \right\} \, . \end{aligned} \tag{100}$$

Here N_f denotes a time-independent normalisation factor, which includes e.g. phase space effects. The decay amplitude describing the transition of the flavour eigenstate B_d in the final state f is denoted by $\mathcal{A}_f$; for the decay of a $\bar{B}_d$ state into f we use the notation $\bar{\mathcal{A}}_f$:

$$\mathcal{A}_f = \langle f | \mathcal{H}_{eff} | B_d \rangle \, , \qquad \bar{\mathcal{A}}_f = \langle f | \mathcal{H}_{eff} | \bar{B}_d \rangle \, . \tag{101}$$

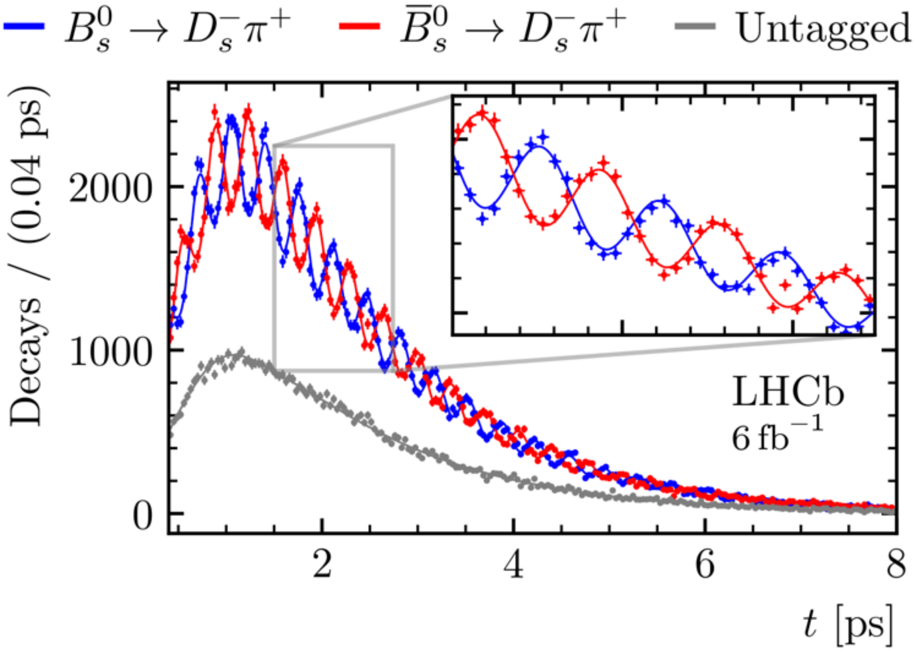
The flavour changing weak quark transitions are described by the effective Hamiltonian described in the previous lecture. The amplitudes $\mathcal{A}_f$ and $\bar{\mathcal{A}}_f$ are typically governed by hadronic effects and they are very difficult to be calculated reliably in theory. Below we will see that CP-symmetries are governed by a single quantity λ_f, which is given by

$$\lambda_f = \frac{q}{p} \frac{\bar{\mathcal{A}}_f}{\mathcal{A}_f} \, . \tag{102}$$

Following the derivation of Eq. (100), the time dependent decay rates of $\Gamma\left[\bar{B}_d(t) \to f\right]$, $\Gamma\left[B_d(t) \to \bar{f}\right]$ and $\Gamma\left[\bar{B}_d(t) \to \bar{f}\right]$ can be derived in a similar way, see e.g. [68, 69] for the analytic expressions.

Remarks:

- The formulae for the time dependent decay rates can be used to extract the observables ΔM_d and $\Delta\Gamma_d$ from experiment, which can then be compared with the theory predictions, see the B_s oscillation plot obtained by LHCb [70].



According to Eq. (84) these observables are related to the matrix elements Γ_{12}^d and M_{12}^d, thus a Standard Model calculation of the mixing observables requires a calculation of the box-diagrams shown above.
In the B_s-system one finds [59] e.g.

$$\Delta M_s^{\text{exp.}} = 17.765(6)\text{ps}^{-1}\,, \qquad \Delta\Gamma_s^{\text{exp.}} = 0.083(5)\text{ps}^{-1}\,, \tag{103}$$

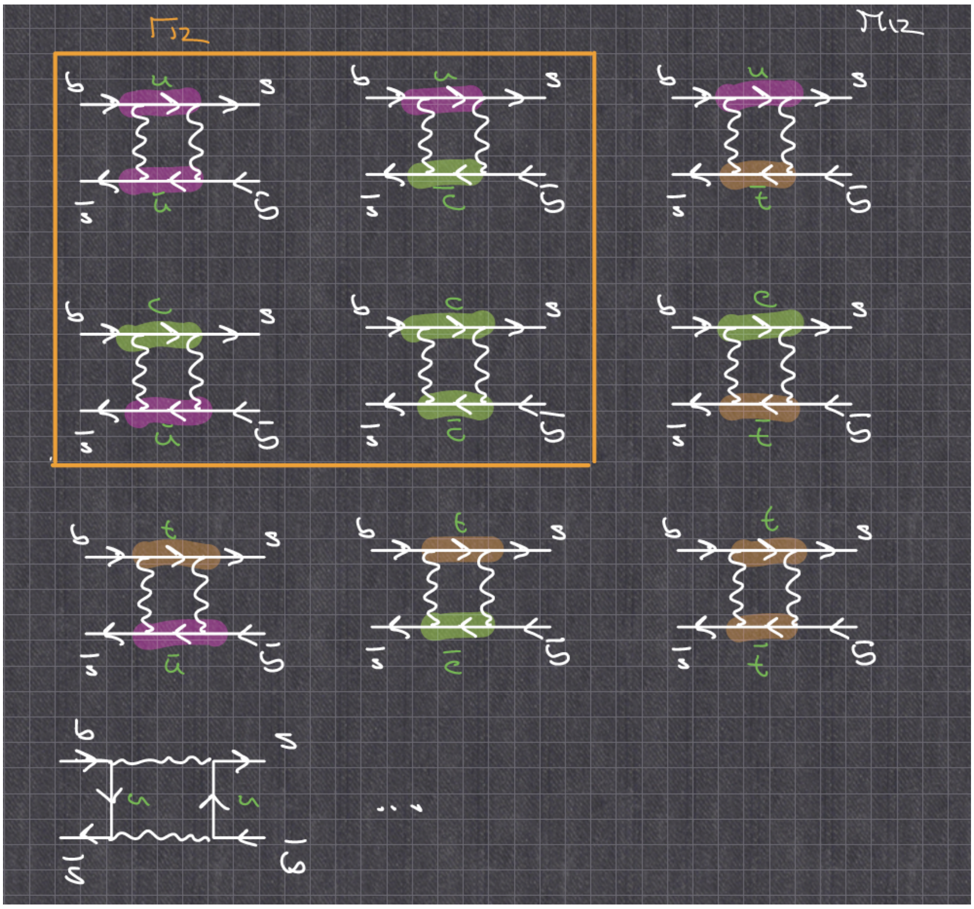
$$\Delta M_s^{\text{th.}} = 18.23(63)\text{ps}^{-1}\,, \qquad \Delta\Gamma_s^{\text{th.}} = 0.091(15)\text{ps}^{-1}\,. \tag{104}$$

- The common pre-factors in the time-dependent decay rates, i.e. N_f and $|\mathcal{A}_f|^2\left(1+|\lambda_f|^2\right)$, typically cancel in CP asymmetries and we do not need to know their value. This is very advantageous because the hadronic quantity $\mathcal{A}_f$ is notoriously difficult to calculate. For the remaining unknown parameter λ_f in some cases additional assumptions can be made, e.g.

  1. In the case of flavour-specific decays, we have $\bar{\mathcal{A}}_f = 0$ and thus $\lambda_f = 0$.
  2. For gold-plated modes we have $f = \bar{f}$ and in addition we consider only one contributing CKM structure in the decay amplitude (in many case this is equivalent to neglect penguin contributions). In that case we arrive at $|\lambda_f| = 1$.

3.3 Theory determination of mixing observables

Mixing arises due to the box diagrams and in leading order there are actually 18 box diagrams contribution to M_{12} and four diagrams contributing to Γ_{12}, see below the B_s-mixing case.



For M_{12}^d we get the following general structure

$$\begin{aligned} M_{12}^d &= \lambda_u^2 F(u,u) + \lambda_u\lambda_c F(u,c) + \lambda_u\lambda_t F(u,t) + \\ & \quad \lambda_c\lambda_u F(c,u) + \lambda_c^2 F(c,c) + \lambda_c\lambda_t F(c,t) + \\ & \quad \lambda_t\lambda_u F(t,u) + \lambda_t\lambda_c F(t,c) + \lambda_t^2 F(t,t)\,, \end{aligned} \tag{105}$$

with the CKM structures $\lambda_q = V_{qd}^* V_{qb}$ for $(q = u, c, t)$ and the loop function $F(q_1, q_2)$ describing internal q_1 and q_2 quarks. For M_{12} we take the off-shell part of the loop diagrams and for Γ_{12} we take the on-shell parts. For B_s mixing we have to use instead the CKM structure $\lambda_q = V_{qs}^* V_{qb}$. In the case of D mixing the internal quarks (u, c, t) are replaced by (d, s, b) and the CKM factor read $\lambda_q = V_{cq}^* V_{uq}$ for $(q = d, s, b)$. In these three systems we get the following CKM hierarchies.

D^0	B_d	B_s
$\lambda_d = V_{cd} V_{ud}^* \propto \lambda$	$\lambda_u = V_{ub} V_{ud}^* \propto \lambda^{3.75}$	$\lambda_u = V_{ub} V_{ud}^* \propto \lambda^{4.75}$
$\lambda_s = V_{cs} V_{us}^* \propto \lambda$	$\lambda_c = V_{cb} V_{cd}^* \propto \lambda^3$	$\lambda_c = V_{cb} V_{cd}^* \propto \lambda^2$
$\lambda_b = V_{cb} V_{ub}^* \propto \lambda^{5.75}$	$\lambda_t = V_{tb} V_{td}^* \propto \lambda^3$	$\lambda_c = V_{tb} V_{td}^* \propto \lambda^2$

(106)

We see a pronounced hierarchy of the CKM elements for D and B_s mixing and a slight one for B_d

mixing. Within the SM we can further use the unitarity of the CKM matrix ($\lambda_u + \lambda_c + \lambda_t = 0 = \lambda_d + \lambda_s + \lambda_b$) to eliminate one of the CKM structure and to make use of numerical hierarchies. For the B_d-system we find

$$\begin{aligned} M_{12}^d &= \lambda_u^2 \left[F(c,c) - 2F(u,c) + F(u,u)\right] \\ &\quad +2\lambda_u\lambda_t \left[F(c,c) - F(u,c) + F(u,t) - F(c,t)\right] \\ &\quad +\lambda_t^2 \left[F(c,c) - 2F(c,t) + F(t,t)\right] \, . \end{aligned} \tag{107}$$

Remarks:

1. In the case of D-mixing Eq. (107) reads

$$\begin{aligned} M_{12}^d &= \lambda_s^2 \left[F(s,s) - 2F(d,s) + F(d,d)\right] \\ &\quad +2\lambda_s\lambda_b \left[F(s,s) - F(d,s) + F(d,b) - F(d,b)\right] \\ &\quad +\lambda_b^2 \left[F(s,s) - 2F(s,b) + F(b,b)\right] \, . \end{aligned} \tag{108}$$

2. GIM cancellations [38]: the general result of a loop calculation looks like

$$F(p,q) = f_0 + f(x_q, x_p) \, , \tag{109}$$

with a constant value f_0 and a mass dependent term $f(x_q, x_p)$ with $x_y = m_y^2/M_W^2$. Thus one finds that f_0 cancels in Eq. (107) due to GIM cancellation - therefore also no renormalisation is necessary if the individual loop diagrams were divergent. If all internal masses would be equal (or zero), M_{12}^q would vanish. Looking at the values of quark masses we find

$$x_u = 7.2 \cdot 10^{-10} \, , \qquad x_d = 3.4 \cdot 10^{-9} \, , \tag{110}$$

$$x_c = 2.5 \cdot 10^{-4} \, , \qquad x_s = 1.3 \cdot 10^{-6} \, , \tag{111}$$

$$x_t = 4 \, , \qquad x_b = 2.7 \cdot 10^{-3} \, . \tag{112}$$

Thus only the top quark has a sizable value of the mass - all other masses are close to be negligible. Thus in the case of B mixing only the last line of Eq. (107) is important, which is also the CKM leading term. In D mixing the same approximation would infer a vanishing result - taking only the effects of the b-quark into account will probably also be a bad approximation, as the last line of Eq. (108) is heavily CKM suppressed. Hence in D-mixing all contributions have to be considered and the result is affected by severe GIM cancellations.

3. Γ_{12} can also be read of Eq. (107) by deleting all terms with a heavy top quark (or b quark in case of D mixing) and by replacing the off-shell function $F(p,q)$ with the on-shell function $F^{OS}(p,q)$

4. In B mixing we find to a very good approximation

$$M_{12}^d = \lambda_t^2 \left[f(t,t) - 2f(c,t) + f(c,c)\right] \propto \lambda_t^2 S(m_t^2/M_W^2) \, , \tag{113}$$

with the Inami-Lim function $S(x)$ [71]:

$$S(x) = \frac{4x - 11x^2 + x^3}{4(1-x)^2} - \frac{3x \ln x}{2(1-x)^2} . \tag{114}$$

Hence we expect

$$\frac{\Delta M_d}{\Delta M_s} = \frac{|V_{td}|^2}{|V_{ts}|^2} = 0.044 , \tag{115}$$

which fits already quite well with the experimental value of 0.03.

Performing the complete calculating of the box diagrams contributing to the last line of Eq. (107) (with internal top and charm quarks) one obtains

$$M_{12}^d = \frac{G_F^2}{12\pi^2} (V_{td}^* V_{tb})^2 M_W^2 S_0(x_t) B_{B_d} f_{B_d}^2 M_{B_d} \hat{\eta}_B . \tag{116}$$

The Inami-Lim function $S_0(x_t = \bar{m}_t^2/M_W^2)$ was discussed above. It results from the box diagram without any gluon corrections. The NLO QCD correction is parameterised by $\hat{\eta}_B \approx 0.84$ [72]. The non-perturbative matrix element of the $\Delta B = 2$ operator

$$O_1 = (\bar{d}b)_{V-A} (\bar{d}b)_{V-A} . \tag{117}$$

is parameterised by the bag parameter B_{B_d} and the decay constant f_{B_d}

$$\langle \bar{B}_d | O_1 | B_d \rangle = \frac{8}{3} f_{B_d}^2 B_{B_d} M_{B_d}^2 . \tag{118}$$

These non-perturbative parameters can be determined with lattice QCD [73] or with the help of QCD sum rules, see e.g. [74]. Current state-of-the-art SM predictions for mixing observables are shown in Eq. (104) and reviewed in [59].

3.4 CP asymmetries

As our last topic we discuss CP violation in hadron decays, where we can distinguish three different origins of CP violation:

1. **CP violation in mixing:** here we consider flavour-specific decays $B \to f$, which means that the decays $\bar{B} \to f$ and $B \to \bar{f}$ are not allowed. Examples for such decays are semi-leptonic B-decays or e.g. $B_s \to D_s^- \pi^+$, which was used for the precision determination of ΔM_s, shown above. We further assume that in these decays no direct CP violation arises, i.e. $A_f = \bar{A}_{\bar{f}}$. Below we will see that this is equivalent to only one contributing CKM structure in the decay amplitude. For such decays we immediately obtain $\lambda_f = 0$, which considerably simplifies the expressions for the time-dependent decay rates in Eq. (100).
   We define the asymmetry for CP violation in mixing as

$$a_{CP}^{\text{mix}} = \frac{\Gamma(\bar{B}_d(t) \to f) - \Gamma(B_d(t) \to \bar{f})}{\Gamma(\bar{B}_d(t) \to f) + \Gamma(B_d(t) \to \bar{f})} . \tag{119}$$

Inserting the expressions for the time-dependent rates (a la Eq. (100)) one finds

$$a_{CP}^{\text{mix}} = a_{fs}^{d} = \Im \left(\frac{\Gamma_{12}^{d}}{M_{12}^{d}} \right) . \tag{120}$$

The flavour-specific CP asymmetry, a_{fs}^{d} has not been measured yet, the current experimental bound reads $a_{fs}^{d,\text{exp.}} = -21(14) \cdot 10^{-4}$, while the SM expectation is $a_{fs}^{d,\text{th.}} = -5.1(0.5) \cdot 10^{-4}$, see [59]. In case of B_s mesons one gets $a_{fs}^{s,\text{exp.}} = -60(280) \cdot 10^{-5}$, while the SM expectation is $a_{fs}^{s,\text{th.}} = 2.2(0.2) \cdot 10^{-5}$. Due to their smallness, the flavour-specific CP asymmetries present excellent opportunities for so-called Null-tests of the SM.

2. **Indirect CP violation or CP violatoin in the interference of mixing and decay:** CP violation was first observed in the decays of Kaons in 1964 as a small effect of the order of several per mille [24], denoted by the quantity ϵ, mentioned in the first lecture. Based on the CKM-mechanism for CP violation, Bigi and Sanda [40] expected in 1981 large (of the order of 50%) CP violating effects in certain B-decays - this was the main motivation for building the B-factories at KEK and SLAC in the 1990s.
Here one investigates the following CP violating asymmetry

$$a_{CP}^{\text{ind}} = \frac{\Gamma(B_d(t) \to f) - \Gamma(B_d(t) \to \bar{f})}{\Gamma(B_d(t) \to f) + \Gamma(B_d(t) \to \bar{f})} . \tag{121}$$

Considering further decays, like $B_d \to J/\psi K_S$, for which $f = \bar{f}$ and which are dominated by a single decay amplitude with the general form

$$\mathcal{A}_f = a \cdot e^{i\phi} \cdot e^{i\theta} , \tag{122}$$

with the modulus of the amplitude a, the strong phase ϕ and the weak phase θ one finds that the parameter λ_f can be simplified as

$$\lambda_f = \frac{q}{p} \frac{\bar{\mathcal{A}}_f}{\mathcal{A}_f} = \frac{q}{p} \frac{\bar{\mathcal{A}}_{\bar{f}}}{\mathcal{A}_f} = \frac{q}{p} \frac{a \cdot e^{i\phi} \cdot e^{-i\theta}}{a \cdot e^{i\phi} \cdot e^{i\theta}} = \frac{q}{p} e^{-2i\theta} . \tag{123}$$

Here all hadronic parameters have canceled! In this case the CP asymmetry a_{CP}^{ind} turns out to be proportional to the sine of twice the angle β of the unitarity triangle. The measurement of a large value of $a_{CP}^{\text{ind}}(B_d \to J/\psi K_S)$ [41,42] confirmed the CKM mechanism and lead to the Nobel Prize for Kobayashi and Maskawa in 2008.
By now the experimental precision in this asymmetry has dramatically improved and currently an uncertainty of β of around $\pm 0.6^\circ$ is quoted [75]. Hence we have to revisit the simplifications made in the theoretical derivation of the asymmetry. In particular, it is well-known that the decay $B_d \to J/\psi K_S$ can also proceed besides the tree-level $b \to c\bar{c}s$-decay via a penguin contribution, leading to two contributing amplitudes

$$\mathcal{A}_f = a \cdot e^{i\phi} \cdot e^{i\theta} + b \cdot e^{i\tilde{\phi}} \cdot e^{i\tilde{\theta}}$$

$$= \quad a \cdot e^{i\phi} \cdot e^{i\theta} \left[1 + \frac{b}{a} \cdot e^{i(\tilde{\phi}-\phi)} \cdot e^{i(\tilde{\theta}-\theta)} \right] , \tag{124}$$

with the modulus of the penguin amplitude b and the strong penguin phase $\tilde{\phi}$ and the weak penguin phase $\tilde{\theta}$. Now the parameter λ_f and the asymmetry get corrections due to the ratio of the penguin and tree amplitude.

$$\lambda_f = \frac{q}{p} e^{-2i\theta} \left[1 + r.... \right] , \tag{125}$$

$$a_{CP}^{\text{ind}} \propto \sin 2\beta \left[1 + r.... \right] , \quad \text{with } r = \frac{b}{a} . \tag{126}$$

This so-called **penguin pollution** is a purely hadronic quantity and therefore very hard to be estimated reliably, but it is expected to be of the order of $\pm 1^\circ$. Hence progress in theory is urgently needed to cope with the increasing experimental precision.

3. **Direct CP violation :** Finally we define the direct CP asymmetry as

$$a_{CP}^{\text{dir}} = \frac{\Gamma(B \to f) - \Gamma(\bar{B} \to \bar{f})}{\Gamma(B \to f) + \Gamma(\bar{B} \to \bar{f})} . \tag{127}$$

Now B can be a neutral or charged B meson. Writing a general decomposition of the decay amplitude for the decay $B \to f$ again as

$$\begin{aligned} \mathcal{A}_f &= a \cdot e^{i\phi} \cdot e^{i\theta} + b \cdot e^{i\tilde{\phi}} \cdot e^{i\tilde{\theta}} \\ &= a \cdot e^{i\phi} \cdot e^{i\theta} \left[1 + r \cdot e^{i(\tilde{\phi}-\phi)} \cdot e^{i(\tilde{\theta}-\theta)} \right] , \end{aligned} \tag{128}$$

with strong phases $\phi, \tilde{\phi}$ and weak phases $\theta, \tilde{\theta}$, we derive the following expression for the direct CP asymmetry

$$a_{CP}^{\text{dir}} \propto r \sin(\tilde{\phi} - \phi) \sin(\tilde{\theta} - \theta) . \tag{129}$$

We find that direct CP violation requires at least two different contributions to the decay amplitude, with different strong and weak phases. Note, that $B_s \to D_s^- \pi^+$ has only one contributing CKM structure, hence no direct CP violation arises, which was used above. Moreover we see, that now the asymmetry is directly proportional to the hadronic ratio r (in the case of indirect CP violation, r was a correction), which is very hard to be determined theoretically. Therefore reliable theory predictions for direct CP asymmetries are very hard to be made, while the theory status of CP violation in mixing and for indirect CP violation is much better under control.

It is interesting to note that recently a large direct CP violation in the decays $D^0 \to \pi^+ \pi^-$ and $D^0 \to K^+ K^-$ has been measured by the LHCb collaboration [76,77], which required values of r being an order of magnitude larger than naively expected. Recent theory progress based on the use of light cone sum rules [47,78] seems to give some evidence for the smallness of r in the Standard Model.

Acknowledgment

I would like to thank the organisers for the invitation to this amazing school, the students for showing intensive interest in both the GIM- and the Gym-mechanism and Maria Laura Piscopo for a careful reading of the notes.

References

[1] S. L. Glashow, Partial Symmetries of Weak Interactions, Nucl. Phys. **22** (1961), 579-588 doi:10.1016/0029-5582(61)90469-2

[2] S. Weinberg, A Model of Leptons, Phys. Rev. Lett. **19** (1967), 1264-1266 doi:10.1103/PhysRevLett.19.1264

[3] A. Salam, Weak and Electromagnetic Interactions Conf. Proc. C **680519** (1968), 367-377 doi:10.1142/9789812795915_0034

[4] P. W. Higgs, Broken Symmetries and the Masses of Gauge Bosons, Phys. Rev. Lett. **13** (1964), 508-509 doi:10.1103/PhysRevLett.13.508

[5] P. W. Higgs, Broken symmetries, massless particles and gauge fields, Phys. Lett. **12** (1964), 132-133 doi:10.1016/0031-9163(64)91136-9

[6] F. Englert and R. Brout, Broken Symmetry and the Mass of Gauge Vector Mesons, Phys. Rev. Lett. **13** (1964), 321-323 doi:10.1103/PhysRevLett.13.321

[7] G. S. Guralnik, C. R. Hagen and T. W. B. Kibble, Global Conservation Laws and Massless Particles, Phys. Rev. Lett. **13** (1964), 585-587 doi:10.1103/PhysRevLett.13.585

[8] T. D. Lee and C. N. Yang, Question of Parity Conservation in Weak Interactions, Phys. Rev. **104** (1956), 254-258 doi:10.1103/PhysRev.104.254

[9] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes and R. P. Hudson, Experimental Test of Parity Conservation in β Decay, Phys. Rev. **105** (1957), 1413-1414 doi:10.1103/PhysRev.105.1413

[10] G. Aad *et al.* [ATLAS], Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Phys. Lett. B **716** (2012), 1-29 doi:10.1016/j.physletb.2012.08.020 [arXiv:1207.7214 [hep-ex]].

[11] S. Chatrchyan *et al.* [CMS], Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC, Phys. Lett. B **716** (2012), 30-61 doi:10.1016/j.physletb.2012.08.021 [arXiv:1207.7235 [hep-ex]].

[12] N. Cabibbo, Unitary Symmetry and Leptonic Decays, Phys. Rev. Lett. **10** (1963), 531-533 doi:10.1103/PhysRevLett.10.531

[13] M. Kobayashi and T. Maskawa, CP Violation in the Renormalizable Theory of Weak Interaction, Prog. Theor. Phys. **49** (1973), 652-657 doi:10.1143/PTP.49.652

[14] J. J. Aubert *et al.* [E598], Experimental Observation of a Heavy Particle J, Phys. Rev. Lett. **33** (1974), 1404-1406 doi:10.1103/PhysRevLett.33.1404

[15] J. E. Augustin *et al.* [SLAC-SP-017], Discovery of a Narrow Resonance in e^+e^- Annihilation, Phys. Rev. Lett. **33** (1974), 1406-1408 doi:10.1103/PhysRevLett.33.1406

[16] L. Wolfenstein, Parametrization of the Kobayashi-Maskawa Matrix, Phys. Rev. Lett. **51** (1983), 1945 doi:10.1103/PhysRevLett.51.1945

[17] J. Charles *et al.* [CKMfitter Group], CP violation and the CKM matrix: Assessing the impact of the asymmetric B factories, Eur. Phys. J. C **41** (2005) no.1, 1-131 doi:10.1140/epjc/s2005-02169-1 [arXiv:hep-ph/0406184 [hep-ph]].

[18] M. Bobrowski, A. Lenz, J. Riedl and J. Rohrwild, How much space is left for a new family of fermions?, Phys. Rev. D **79** (2009), 113006 doi:10.1103/PhysRevD.79.113006 [arXiv:0902.4883 [hep-ph]].

[19] O. Eberhardt, A. Lenz and J. Rohrwild, Less space for a new family of fermions, Phys. Rev. D **82** (2010), 095006 doi:10.1103/PhysRevD.82.095006 [arXiv:1005.3505 [hep-ph]].

[20] A. Djouadi and A. Lenz, Sealing the fate of a fourth generation of fermions, Phys. Lett. B **715** (2012), 310-314 doi:10.1016/j.physletb.2012.07.060 [arXiv:1204.1252 [hep-ph]].

[21] O. Eberhardt, G. Herbert, H. Lacker, A. Lenz, A. Menzel, U. Nierste and M. Wiebusch, Impact of a Higgs boson at a mass of 126 GeV on the standard model with three and four fermion generations, Phys. Rev. Lett. **109** (2012), 241802 doi:10.1103/PhysRevLett.109.241802 [arXiv:1209.1101 [hep-ph]].

[22] P. A. R. Ade *et al.* [Planck], Planck 2013 results. VI. High Frequency Instrument data processing, Astron. Astrophys. **571** (2014), A6 doi:10.1051/0004-6361/201321570 [arXiv:1303.5067 [astro-ph.CO]].

[23] A. D. Sakharov, Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe, Pisma Zh. Eksp. Teor. Fiz. **5** (1967), 32-35 doi:10.1070/PU1991v034n05ABEH002497

[24] J. H. Christenson, J. W. Cronin, V. L. Fitch and R. Turlay, Evidence for the 2π Decay of the K_2^0 Meson, Phys. Rev. Lett. **13** (1964), 138-140 doi:10.1103/PhysRevLett.13.138

[25] O. Atkinson, M. Black, A. Lenz, A. Rusov and J. Wynne, Cornering the Two Higgs Doublet Model Type II, JHEP **04** (2022), 172 doi:10.1007/JHEP04(2022)172 [arXiv:2107.05650 [hep-ph]].

[26] N. S. Manton, Topology in the Weinberg-Salam Theory, Phys. Rev. D **28** (1983), 2019 doi:10.1103/PhysRevD.28.2019

[27] F. R. Klinkhamer and N. S. Manton, A Saddle Point Solution in the Weinberg-Salam Theory, Phys. Rev. D **30** (1984), 2212 doi:10.1103/PhysRevD.30.2212

[28] C. Jarlskog, Commutator of the Quark Mass Matrices in the Standard Electroweak Model and a Measure of Maximal CP Nonconservation, Phys. Rev. Lett. **55** (1985), 1039 doi:10.1103/PhysRevLett.55.1039

[29] M. E. Shaposhnikov, Possible Appearance of the Baryon Asymmetry of the Universe in an Electroweak Theory, JETP Lett. **44** (1986), 465-468

[30] G. Alonso-Álvarez, G. Elor and M. Escudero, Collider signals of baryogenesis and dark matter from B mesons: A roadmap to discovery, Phys. Rev. D **104** (2021) no.3, 035028 doi:10.1103/PhysRevD.104.035028 [arXiv:2101.02706 [hep-ph]].

[31] K. Kajantie, M. Laine, K. Rummukainen and M. E. Shaposhnikov, Is there a hot electroweak phase transition at $m_H \gtrsim m_W$?, Phys. Rev. Lett. **77** (1996), 2887-2890 doi:10.1103/PhysRevLett.77.2887 [arXiv:hep-ph/9605288 [hep-ph]].

[32] K. Rummukainen, M. Tsypin, K. Kajantie, M. Laine and M. E. Shaposhnikov, The Universality class of the electroweak theory, Nucl. Phys. B **532** (1998), 283-314 doi:10.1016/S0550-3213(98)00494-5 [arXiv:hep-lat/9805013 [hep-lat]].

[33] F. Csikor, Z. Fodor and J. Heitger, Endpoint of the hot electroweak phase transition, Phys. Rev. Lett. **82** (1999), 21-24 doi:10.1103/PhysRevLett.82.21 [arXiv:hep-ph/9809291 [hep-ph]]. [34]

[34] Y. Aoki, F. Csikor, Z. Fodor and A. Ukawa, The Endpoint of the first order phase transition of the SU(2) gauge Higgs model on a four-dimensional isotropic lattice, Phys. Rev. D **60** (1999), 013001 doi:10.1103/PhysRevD.60.013001 [arXiv:hep-lat/9901021 [hep-lat]].

[35] O. Atkinson, M. Black, C. Englert, A. Lenz, A. Rusov and J. Wynne, The flavourful present and future of 2HDMs at the collider energy frontier, JHEP **11** (2022), 139 doi:10.1007/JHEP11(2022)139 [arXiv:2202.08807 [hep-ph]].

[36] O. Atkinson, M. Black, C. Englert, A. Lenz and A. Rusov, MUonE, muon g-2 and electroweak precision constraints within 2HDMs, Phys. Rev. D **106** (2022) no.11, 115031 doi:10.1103/PhysRevD.106.115031 [arXiv:2207.02789 [hep-ph]].

[37] D. E. Morrissey and M. J. Ramsey-Musolf, Electroweak baryogenesis, New J. Phys. **14** (2012), 125003 doi:10.1088/1367-2630/14/12/125003 [arXiv:1206.2942 [hep-ph]].

[38] S. L. Glashow, J. Iliopoulos and L. Maiani, Weak Interactions with Lepton-Hadron Symmetry, Phys. Rev. D **2** (1970), 1285-1292 doi:10.1103/PhysRevD.2.1285

[39] D. Abbaneo, A. Ali, P. Amaral, V. Andreev, M. Artuso, E. Barberio, M. Battaglia, C. Bauer, D. Becirevic and M. Beneke, *et al.* The CKM matrix and the unitarity triangle. Workshop, CERN, Geneva, Switzerland, 13-16 Feb 2002: Proceedings, doi:10.5170/CERN-2003-002-corr [arXiv:hep-ph/0304132 [hep-ph]].

[40] I. I. Y. Bigi and A. I. Sanda, Notes on the Observability of CP Violations in B Decays, Nucl. Phys. B **193** (1981), 85-108 doi:10.1016/0550-3213(81)90519-8

[41] K. Abe *et al.* [Belle], Observation of large CP violation in the neutral B meson system, Phys. Rev. Lett. **87** (2001), 091802 doi:10.1103/PhysRevLett.87.091802 [arXiv:hep-ex/0107061 [hep-ex]].

[42] B. Aubert *et al.* [BaBar], Observation of CP violation in the B^0 meson system, Phys. Rev. Lett. **87** (2001), 091801 doi:10.1103/PhysRevLett.87.091801 [arXiv:hep-ex/0107013 [hep-ex]]. Copy to ClipboardDownload

[43] J. Brod and J. Zupan, The ultimate theoretical error on γ from $B \to DK$ decays, JHEP **01** (2014), 051 doi:10.1007/JHEP01(2014)051 [arXiv:1308.5663 [hep-ph]].

[44] J. Brod, A. Lenz, G. Tetlalmatzi-Xolocotzi and M. Wiebusch, New physics effects in tree-level decays and the precision in the determination of the quark mixing angle γ, Phys. Rev. D **92** (2015) no.3, 033002 doi:10.1103/PhysRevD.92.033002 [arXiv:1412.1446 [hep-ph]].

[45] A. Lenz and G. Tetlalmatzi-Xolocotzi, Model-independent bounds on new physics effects in non-leptonic tree-level decays of B-mesons, JHEP **07** (2020), 177 doi:10.1007/JHEP07(2020)177 [arXiv:1912.07621 [hep-ph]].

[46] M. Bona *et al.* [UTfit], The Unitarity Triangle Fit in the Standard Model and Hadronic Parameters from Lattice QCD: A Reappraisal after the Measurements of ΔM_s and $BR(B \to \tau\nu_\tau)$, JHEP **10** (2006), 081 doi:10.1088/1126-6708/2006/10/081 [arXiv:hep-ph/0606167 [hep-ph]].

[47] A. Lenz, M. L. Piscopo and A. V. Rusov, Two body non-leptonic D^0 decays from LCSR and implications for$\Delta a_{\mathrm{CP}}^{\mathrm{dir}}$, JHEP **03** (2024), 151 doi:10.1007/JHEP03(2024)151 [arXiv:2312.13245 [hep-ph]].

[48] L. Michel, Interaction between four half spin particles and the decay of the μ meson, Proc. Phys. Soc. A **63** (1950), 514-531 doi:10.1088/0370-1298/63/5/311

[49] R. L. Workman *et al.* [Particle Data Group], Review of Particle Physics, PTEP **2022** (2022), 083C01 doi:10.1093/ptep/ptac097

[50] R. E. Behrends, R. J. Finkelstein and A. Sirlin, Radiative corrections to decay processes, Phys. Rev. **101** (1956), 866-873 doi:10.1103/PhysRev.101.866

[51] T. Kinoshita and A. Sirlin, Radiative corrections to Fermi interactions, Phys. Rev. **113** (1959), 1652-1660 doi:10.1103/PhysRev.113.1652

[52] T. van Ritbergen and R. G. Stuart, Complete two loop quantum electrodynamic contributions to the muon lifetime in the Fermi model, Phys. Rev. Lett. **82** (1999), 488-491 doi:10.1103/PhysRevLett.82.488 [arXiv:hep-ph/9808283 [hep-ph]].

[53] A. Sirlin and A. Ferroglia, Radiative Corrections in Precision Electroweak Physics: a Historical Perspective, Rev. Mod. Phys. **85** (2013) no.1, 263-297 doi:10.1103/RevModPhys.85.263 [arXiv:1210.5296 [hep-ph]].

[54] P. A. Baikov, K. G. Chetyrkin and J. H. Kuhn, Order α_s^4 QCD Corrections to Z and tau Decays, Phys. Rev. Lett. **101** (2008), 012002 doi:10.1103/PhysRevLett.101.012002 [arXiv:0801.1821 [hep-ph]].

[55] A. Pich, Precision Tau Physics, Prog. Part. Nucl. Phys. **75** (2014), 41-85 doi:10.1016/j.ppnp.2013.11.002 [arXiv:1310.7922 [hep-ph]].

[56] S. Bethke, Look how it runs!, [arXiv:2310.01111 [hep-ph]].

[57] M. L. Piscopo, Higher order corrections to the lifetime of heavy hadrons, doi:10.25819/ubsi/10024 [arXiv:2112.03137 [hep-ph]].

[58] A. Lenz, Lifetimes and heavy quark expansion, Int. J. Mod. Phys. A **30** (2015) no.10, 1543005 doi:10.1142/S0217751X15430058 [arXiv:1405.3601 [hep-ph]].

[59] J. Albrecht, F. Bernlochner, A. Lenz and A. Rusov, Lifetimes of b-hadrons and mixing of neutral B-mesons: theoretical and experimental status, Eur. Phys. J. ST **233** (2024) no.2, 359-390 doi:10.1140/epjs/s11734-024-01124-3 [arXiv:2402.04224 [hep-ph]].

[60] A. J. Buras, Weak Hamiltonian, CP violation and rare decays," [arXiv:hep-ph/9806471 [hep-ph]].

[61] G. Buchalla, A. J. Buras and M. E. Lautenbacher, Weak decays beyond leading logarithms, Rev. Mod. Phys. **68** (1996), 1125-1144 doi:10.1103/RevModPhys.68.1125 [arXiv:hep-ph/9512380 [hep-ph]].

[62] A. Lenz, Heavy Flavour Physics and Effective Field Theories, Lecture notes, weblink: https://tp.nt.uni-siegen.de/wp-content/uploads/2024/02/Lecture_Flav_2021.pdf.

[63] G. Altarelli and L. Maiani, Phys. Lett. B **52** (1974), 351-354 doi:10.1016/0370-2693(74)90060-4

[64] M. Gorbahn and U. Haisch, Effective Hamiltonian for non-leptonic $|\Delta F| = 1$ decays at NNLO in QCD, Nucl. Phys. B **713** (2005), 291-332 doi:10.1016/j.nuclphysb.2005.01.047 [arXiv:hep-ph/0411071 [hep-ph]].

[65] M. A. Shifman, Foreword to ITEP lectures in particle physics, [arXiv:hep-ph/9510397 [hep-ph]].

[66] J. Albrecht, D. van Dyk and C. Langenbruch, Flavour anomalies in heavy quark decays, Prog. Part. Nucl. Phys. **120** (2021), 103885 doi:10.1016/j.ppnp.2021.103885 [arXiv:2107.04822 [hep-ex]].

[67] A. Lenz and G. Wilkinson, Mixing and CP Violation in the Charm System, Ann. Rev. Nucl. Part. Sci. **71** (2021), 59-85 doi:10.1146/annurev-nucl-102419-124613 [arXiv:2011.04443 [hep-ph]].

[68] M. Artuso, G. Borissov and A. Lenz, CP violation in the B_s^0 system, Rev. Mod. Phys. **88** (2016) no.4, 045002 doi:10.1103/RevModPhys.88.045002 [arXiv:1511.09466 [hep-ph]].

[69] K. Anikeev, D. Atwood, F. Azfar, S. Bailey, C. W. Bauer, W. Bell, G. Bodwin, E. Braaten, G. Burdman and J. N. Butler, *et al.* B physics at the Tevatron: Run II and beyond, [arXiv:hep-ph/0201071 [hep-ph]].

[70] R. Aaij *et al.* [LHCb], Precise determination of the $B^0_{\rm s}$–$\overline{B}{}^0_{\rm s}$ oscillation frequency, Nature Phys. **18** (2022) no.1, 1-5 doi:10.1038/s41567-021-01394-x [arXiv:2104.04421 [hep-ex]].

[71] T. Inami and C. S. Lim, Effects of Superheavy Quarks and Leptons in Low-Energy Weak Processes $K_L \to \mu^+\mu^-$, $K^+ \to \pi^+\nu\bar{\nu}$, and $K^0 \leftrightarrow \bar{K}^0$, Prog. Theor. Phys. **65** (1981), 297 [erratum: Prog. Theor. Phys. **65** (1981), 1772] doi:10.1143/PTP.65.297

[72] A. J. Buras, M. Jamin and P. H. Weisz, Leading and Next-to-leading QCD Corrections to ϵ Parameter and $B^0 - \bar{B}^0$ Mixing in the Presence of a Heavy Top Quark, Nucl. Phys. B **347** (1990), 491-536 doi:10.1016/0550-3213(90)90373-L

[73] Y. Aoki *et al.* [Flavour Lattice Averaging Group (FLAG)], FLAG Review 2021, Eur. Phys. J. C **82** (2022) no.10, 869 doi:10.1140/epjc/s10052-022-10536-1 [arXiv:2111.09849 [hep-lat]].

[74] D. King, A. Lenz and T. Rauh, B$_s$ mixing observables and |V$_{td}$/V$_{ts}$| from sum rules, JHEP **05** (2019), 034 doi:10.1007/JHEP05(2019)034 [arXiv:1904.00940 [hep-ph]].

[75] R. Aaij *et al.* [LHCb], Measurement of CP Violation in B0$\to\psi(\to\ell$+ℓ-)KS0($\to\pi$+π-) Decays, Phys. Rev. Lett. **132** (2024) no.2, 021801 doi:10.1103/PhysRevLett.132.021801 [arXiv:2309.09728 [hep-ex]].

[76] R. Aaij *et al.* [LHCb], Observation of CP Violation in Charm Decays, Phys. Rev. Lett. **122** (2019) no.21, 211803 doi:10.1103/PhysRevLett.122.211803 [arXiv:1903.08726 [hep-ex]].

[77] R. Aaij *et al.* [LHCb], Measurement of the Time-Integrated CP Asymmetry in D0$\to$K-K+ Decays, Phys. Rev. Lett. **131** (2023) no.9, 091802 doi:10.1103/PhysRevLett.131.091802 [arXiv:2209.03179 [hep-ex]].

[78] A. Lenz, M. L. Piscopo and A. V. Rusov, Towards a SM prediction for CP violation in charm, [arXiv:2403.02267 [hep-ph]].

Machine learning

Jan Kieseler[a]

[a]KIT, Karlsruhe, Germany

Advanced machine learning techniques have become ubiquitous: from computer vision algorithms found on a plethora of small devices such as cameras or smartphones to the recent rise of tremendously powerful large language models. Also in high energy particle physics, these techniques have become essential and have led to a significant increase in physics reach, from simple feed-forward algorithms used to distinguish signal and background processes to more complex neural networks that utilise the underlying physics structure of the data. This section will cover the basics of neural networks and their training and will then discuss examples of the building blocks that make up modern machine learning algorithms, aiming to provide a tool box for their further application in physics analyses.

These lecture notes give a short overview of the basic principles of machine learning, exclusively focusing on deep learning. They also cover neural network building blocks adapting to the structure of the data that should be processed by the algorithms and discuss examples of applications in high energy physics. There is already a large amount of literature covering these topics and reliable and well curated material on the internet, in particular for the basic principles, which is why the overview here is kept short and aims to give an intuitive understanding of the ideas behind the covered machine learning techniques. A very concise, yet rather comprehensive summary of the basic principles and neural network building blocks can be found for example in Ref. [1], also containing references to the relevant publications. In principle, the material that is worth covering in this context would correspond to at least one full course. Therefore, only a few examples of relevant methods and techniques are discussed here with the intent that they could serve as a stepping stone towards a more in-depth study of the subject or as a starting point for applications of deep neural networks to physics studies. For the latter, knowing in particular what it means to exploit the structure of the data in the architecture of the neural network is useful, which is why this part will receive the most focus.

This chapter should be cited as: Machine learning, Jan Kieseler, DOI: 10.23730/CYRSP-2025-009.189,
in: Proceedings of the 2023 European School of High-Energy Physics,
CERN Yellow Reports: School Proceedings, CERN-2025-009, DOI: 10.23730/CYRSP-2025-009, p.189.

1 Basic principles

The most basic building block of a neural network is a dense neural network layer. The operation a dense layer ϕ performs given the input vector x can be described by a weight matrix ω, a bias vector b and an activation function θ:

$$h = \phi(x) = \theta(\omega\, x + b), \tag{1}$$

where h is the output vector. The parameters of ω as well as b are free parameters and are learned during the neural network training through gradient descent, covered in more detail later, where the learnable parameters are updated such that the difference between the neural network output and the desired *true* output is minimised. The weight matrix ω can change the dimensionality of the output with respect to the input. A deep feed-forward neural network, also referred to as multi-layer perceptron (MLP), can be built by stacking these dense neural network layers, where each layer k passes on its output to the next $(k+1)$, in other words:

$$h^{(k+1)}(h^{(k)}) = \theta(\omega_{k+1} h^{(k)} + b_{k+1}). \tag{2}$$

This equation recursively defines an MLP, also depicted in Figure 1, where the information is passed from an input layer through hidden layers to the output layer. At each layer, an element of the (hidden) representation is also often referred to as *node*. By learning the free parameters of that model, an MLP can act as a universal function approximator.

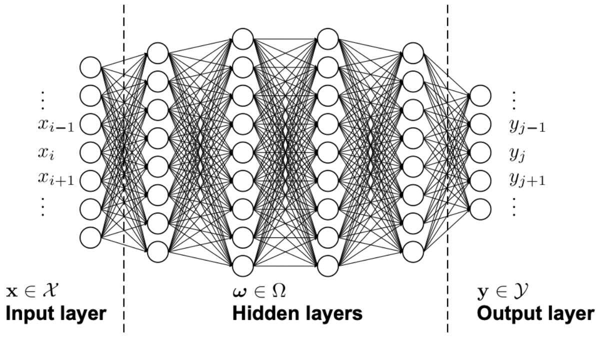

x_{i-1}
x_i
x_{i+1}
y_{j-1}
y_j
y_{j+1}
$\mathbf{x} \in \mathcal{X}$
Input layer
$\boldsymbol{\omega} \in \Omega$
Hidden layers
$\mathbf{y} \in \mathcal{Y}$
Output layer

Fig. 1: A sketch of a feed-forward neural network, or multi-layer perceptron. Each node is illustrated by a circle and weights are illustrated by connections. Activation functions and bias vectors are omitted.

An activation function θ appearing in each layer is needed to enable the neural network to describe non-linear dependencies. This can easily be verified if θ is omitted:

$$h^{(k+1)}(h^{(k)}) = \omega_{k+1} h^{(k)} + b_{k+1} = \omega_{k+1}\left(\omega_k h^{(k-1)} + b_k\right) + b_{k+1}. \tag{3}$$

Setting $\omega = \omega_{k+1}\omega_k$ and $b = \omega_{k+1} b_k + b_{k+1}$, the operation of network layer k and $k+1$ can be reduced to one operation. Recursively applied to a neural network of arbitrary depth, this would allow to represent a neural network of arbitrary depth by a single linear operation, limiting the set of functions it can approximate, the expressivity, significantly.

While the only conceptual requirement for the activation function in a hidden layer is that it is non-linear, in practice, there are some restrictions: it should provide numerically stable output and a numerically stable gradient. As it will be applied many times within the neural network it should also be computationally simple and it should not significantly scale the hidden representations, which could lead to *vanishing* or *exploding* outputs and gradients. Therefore, typical activation functions are applied element wise to the vectors and have a derivative of about one at zero. Two often used examples are tanh, RELU and ELU:

$$\mathrm{RELU}(x) = \max(x, 0) \tag{4}$$

$$\mathrm{ELU}(x) = \begin{array}{ll} e^x - 1 & 0 > x \\ x & 0 \leq x \end{array} \tag{5}$$

In most cases, the exact choice of the hidden-layer activation functions does not have a large impact on the final result. A more comprehensive survey can be found e.g. in Ref. [2].

1.1 Training

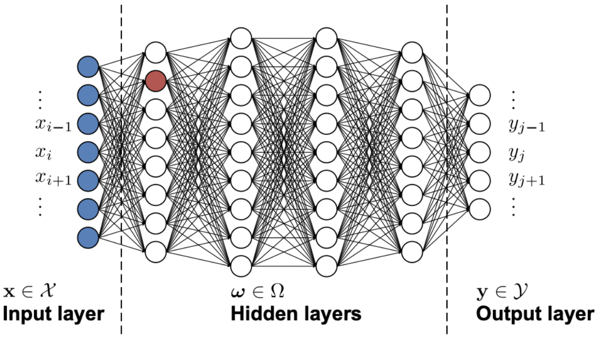


Fig. 2: Activation of a single node in the neural network.

Keeping variance of the inputs, the hidden layers, and the outputs within reasonable constraints is useful for a fast convergence of the training process. This is very similar to practical considerations when performing fits of a functional form to data points (e.g. discussed in Chapter **Practical Statistics**), where also well chosen start parameters as well as reasonably defined parameter ranges help the fit to converge. Therefore, also the input data x is typically normalised, such that the values have approximately a mean of zero and a variance of one. Assuming N inputs that are uncorrelated and normal distributed, the distribution at the red node in Figure 2 would also correspond to a normal distribution, with a variance of N. To avoid that the variance increases in a similar manner with each layer, the weights ω_1 of the first hidden layer can be initialised also following a normal distribution, but scaled by $1/\sqrt{N}$, such that the variance remains one. This choice is referred to as Glorot initialisation [3]. Since also the activation function is applied before feeding the output to the next layer, it is often beneficial to chose initialisation and activation functions that preserve this property, at least to a good approximation. For example, Glorot can be paired with a tanh activation function, while ELU and RELU can be paired with He initialisation [4].

To train the parameters of the neural network, a cost (or loss) function needs to be defined. This function quantifies how well the neural network approximates the desired output, given a set of weights. An example is a classic linear regression, where the parameters a and b of a function $f(x) = ax + b$ are fit to data points. As discussed in Chapter **Practical Statistics**, this can be done e.g. by using the least square method or through a χ^2 minimisation. In the context of neural networks, the mean of the least squares or the χ^2 define a loss function for fitting the parameters a and b. In both cases, the value of the function $f(x)$ is compared to the true value of y for each point and the (weighted) mean difference is minimised. The training of a neural network follows the exact same principle, but with orders of magnitude more free parameters and typically with a larger set of training points (or data set). One of the most common used loss functions for regression tasks is the mean-squared error (MSE) loss, in this case evaluated for the neural network Φ:

$$\min 1/N \sum_{i}^{N} \left((\Phi(\omega,\, x_i) - y_i)^2 \right) = \min \mathrm{MSE}(\Phi(\omega,\, x), y). \tag{6}$$

Here, x, y, and ω represent all points or all weights respectively. In this case, the neural network should not have a restricted output range, which is why the output layer should have no activation (also referred to as linear activation). In this context, it is important to make the connection to the origin of choosing the MSE loss or a χ^2: the network output is expected to follow a Gaussian likelihood, and therefore minimising the loss function corresponds to minimising the negative log-likelihood that is representative of the problem.

Another prominent task for a neural network is classification, e.g. for distinguishing cat pictures from dog pictures, or signal events from background events. This is a binary classification problem, where the probability for a sample to be identified by the neural network corresponds to a Bernoulli process. With $\hat{y} =: \Phi(\omega, x)$, the probability for a single sample to be identified by the neural network becomes:

$$P(\hat{y}, y) = \hat{y}^y (1 - \hat{y})^{1-y} \tag{7}$$

There, the likelihood for N independent samples factorises as

$$\Pi_{l=1}^{N} (\hat{y}^{(l)})^{y^{(l)}} (1 - \hat{y}^{(l)})^{(1-y^{(l)})} \tag{8}$$

and the negative logarithm becomes:

$$\sum_{l}^{N} \left(y^{(l)} \log(\hat{y}^{(l)}) + (1 - y^{(l)}) \log(1 - \hat{y}^{(l)}) \right), \tag{9}$$

which is the expression for the binary cross entropy loss used to train binary classification neural networks. Since the probability is strictly defined between zero and one, the activation function for the output layer of the neural network also needs to be bound to the same interval. For this purpose, a sigmoid activation is used for the output layer:

$$\theta_{\text{sigmoid}}(x) = 1/(1 + e^{-x}). \tag{10}$$

In general, this underlines that the choice of the loss function is defined by the output distribution the network is expected to have. In many cases it can be beneficial to either adapt the loss function or redefine the output of the neural network. One example would be if a neural network should be trained to refine the momentum estimate of a reconstructed jet. Here it can be more beneficial to learn a correction to the unrefined momentum estimate, approximately Gaussian distributed with a mean around one, than to learn to estimate the corrected energy directly, realistically covering an interval that spans two orders of magnitude.

Finally, the training of the neural network is performed using (variants of) gradient descent. Gradient descent is a well established step-wise robust minimisation method, where the parameters of the neural network (or a function) are updated using the gradient of the loss function with respect to the parameters:

$$\omega^{(s+1)} = \omega^{(s)} - \eta \nabla_{\omega^{(s)}} L\left(\Phi(\omega^{(s)},\, x),\, y\right). \tag{11}$$

Here, L is the loss function, s is an update step, $\eta > 0$ is a parameter referred to as learning rate, and x, y, and ω stand for the whole set of training data points or all weights, respectively. The procedure is repeated until a stopping criterion is reached and the loss value does not improve significantly anymore:

$$L\left(\Phi(\omega^{(s)},\, x),\, y\right) - L\left(\Phi(\omega^{(s+1)},\, x),\, y\right) < \epsilon. \tag{12}$$

While gradient descent is only one of many techniques that can be used in principle to minimise the loss function, it is central to all modern deep learning tasks. One of the main reasons is that often the data set used for training, as well as the neural network itself are too large to perform gradient descent on the full data set. In this case, gradient descent naturally extends to stochastic gradient descent on (mini) batches of the data, where only a part of the data set $(\{x\}_s,\, \{y\}_s)$ is processed at each step s:

$$\omega^{(s+1)} = \omega^{(s)} - \eta \nabla_{\omega^{(s)}} L\left(\Phi(\omega,\, \{x\}_s),\, \{y\}_s\right) \tag{13}$$

This reduces the computational burden and makes the training of large models feasible. However, it has implications on the stopping criterion and the learning rate as the procedure introduces additional noise into the system. In practice that means that the learning rate needs to be decreased the smaller the batches become, and that the stopping criterion needs to be tuned to account for statistical fluctuations due to the batch size and fluctuations in the gradients from batch to batch. On the other hand, this additional noise can help to find a minimum. Considering that deep neural networks have thousands to billions of parameters, it is not guaranteed that there is always a clear path to one global minimum. More concretely, it is not guaranteed that the optimisation problem is strictly convex. In these situations it has been shown empirically that some additional noise can help to avoid local minima and increases the chance to converge towards a global minimum (or a valley in parameter space that minimises the loss function) [5]. Typically, the algorithms that performs updates of the weights of the neural network are referred to as optimisers.

One disadvantage of optimisers using stochastic gradient descent alone is that they need many steps to converge as parameters get updated only in small steps, and each step is independent of the previous ones. This problem can be solved by introducing momentum, which can be interpreted very closely to the physical momentum when the loss landscape is interpreted as a physical potential. In this interpretation, the gradient descent introduces a force and the momentum mechanism adds a mass as well as friction to the system. The resulting behaviour is illustrated in Figure 3.

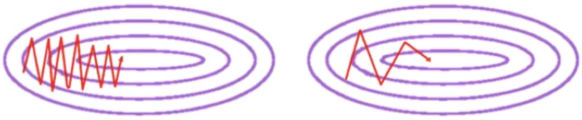


Fig. 3: Illustration of stochastic gradient descent, without momentum (left) and with momentum (right). Figure taken from [5].

Concretely, this means, we introduce a momentum ν at each step s, where

$$v^{(s)} = \alpha v^{(s-1)} - \eta \nabla_{\omega^{(s)}} L, \tag{14}$$

and a low value for the parameter α corresponds to large friction. Then, the parameter update becomes:

$$\omega^{(s+1)} = \omega^{(s)} + v^{(s)}. \tag{15}$$

Many modern optimisers implement different variants of momentum to reduce the number of steps needed for convergence. An overview can be found in the literature (e.g. in Refs. [1,5]) and visualisations are available online, e.g. in Ref. [6]. A key ingredient to gradient-based optimisation in all cases is access to the gradients. While in principle the gradients could be calculated numerically at each step, the amount of parameters in a typical deep neural network makes this inefficient and not feasible. Instead, gradients are calculated analytically in modern machine-learning frameworks, but typically without the explicit need of a user to do define their calculation manually. This is referred to as auto-differentiation and has contributed significantly to the success of deep learning approaches [7]. In auto-differentiation, each operation (e.g. the matrix multiplication in a dense neural network layer) is also assigned to an analytic gradient. Taking for example a very simple neural network

$$\Phi(\omega, x) = \tanh(\omega x) \tag{16}$$

and a loss function

$$L = (\Phi - y)^2, \tag{17}$$

Considering each operation separately, we also define $a = \omega x$. Then gradients for ω can be calculated based on the chain rule as:

$$\frac{\partial L}{\partial \omega} = \frac{\partial a}{\partial \omega}\frac{\partial \Phi}{\partial a}\frac{\partial L}{\partial \Phi} = [x] \cdot \left[1 - \tanh^2(a)\right] \cdot [2(\Phi - y)] \tag{18}$$

At a certain step of the optimisation, this analytic expression only needs to be evaluated for a certain numerical choice for x, y, and ω. This can easily be done by first performing a *forward pass* in which the neural network is evaluated, while saving the individual outputs for each operation: a, and Φ. Now the gradient with respect to ω can be calculated by evaluating the above expression, using those intermediate results. Moreover, it can be seen from the equation that even if x were an output of a previous neural-network layer, one could calculate the gradient for ω first when going backwards, and then use the numerical output of that to calculate the gradients for the weights in this hypothetical previous layer. This process is called the *backpropagation*. It makes calculating gradients in large and complex neural networks computationally feasible and offers a simple way of calculating gradients by attaching analytical derivatives to fundamental operations only [7].

For a training to converge within a finite (best short) time, many parameters such as the learning rates, but also momentum parameters need to be tuned. With modern optimisers, the momentum parameters are often chosen well and require less tuning. The learning rate, however, depends strongly on the model complexity, the batch size, the loss function, and the training data. In practice, it needs to be tuned for each model by observing the loss value during the training. The training is typically performed in multiple iterations over the full training data set, referred to as an epoch.

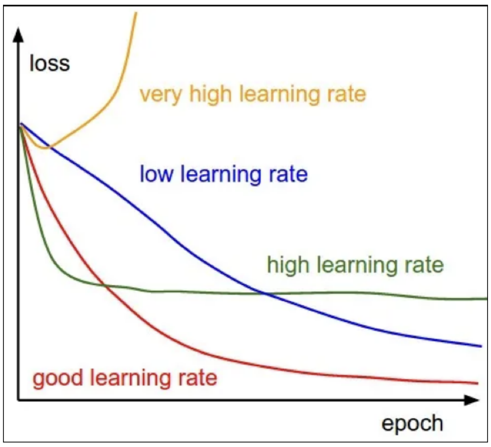


Fig. 4: Sketch of the behaviour of the loss during training for different learning rates. Figure taken from [8].

As illustrated in Figure 4, a too low learning rate can increase the training time significantly (and can also lead to the optimisation process getting stuck in local minima), while a too high learning rate may lead to missing the minimum or even no convergence at all. These effects can be intuitively understood with the help of Figure 5.

Since there is no generally applicable "best" learning rate, it is advisable to test a few learning rates in practice to gauge e.g. which learning rate is too high and at which rate the training stagnates. Typically, learning rates do not exceed 10^{-2} and are often in the range of 10^{-4}, depending on the neural network architecture. Also the best choices for parameters such as the number of nodes at each layer or the depth of the neural network need to be tuned by hand or by using other optimisation algorithms outside of the training procedure. These parameters are also called *hyper parameters*.

As mentioned before, deep neural networks can serve as universal function approximators. In case the neural network has sufficient expressivity (e.g. by having a large amount of free parameters), it can in principle learn each individual sample in the data set it is trained with. In general, this is not desirable, since it can (but curiously does not have to) imply that the network does not generalise, referred to as overtraining or overfitting. For the HEP context, that would mean it would learn to identify e.g.

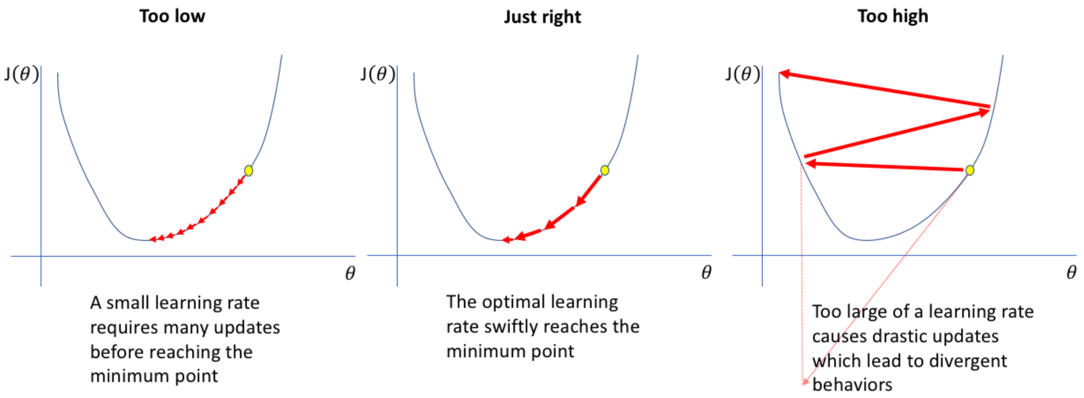

Too low
Just right
Too high
J(θ)
J(θ)
J(θ)
θ
θ
θ
A small learning rate requires many updates before reaching the minimum point
The optimal learning rate swiftly reaches the minimum point
Too large of a learning rate causes drastic updates which lead to divergent behaviors

Fig. 5: Sketch of the behaviour of the loss during training for different learning rates. The loss value is represented by $J(\theta)$, and the parameter to be optimised by θ. Figure taken from [9].

individual collision events, but would not learn generic properties of the underlying physics process. If such a network is then confronted with new data, it would not give the desired result. To identify such behaviour, one usually defines a *validation* data set. In most cases, it is split off from the training data set and consists of samples that are never used in the training. By monitoring the loss in this validation data set during training, overtraining can be detected by an increase in the loss for the validation data set while the loss for the training data set still reduces. There are multiple ways to reduce overtraining, such as different regularisation techniques that also concisely described in Ref. [1] and others. Here, I would like to focus on the simplest approach: more data samples per neural network weight. This means either reducing the network complexity or creating more training data. While the latter is a large challenge in computer science, where many data sets rely on humans to assign a truth label (e.g. if a picture contains a cat or a dog), the situation in HEP is usually different. Here, we would usually train the models on simulated events that we can produce with very little effort in comparison. Therefore it is generally easier to produce a larger data set than studying different regularisation schemes involving a large set of additional parameters to be tuned. The same should be mentioned with respect to data augmentation and other techniques that address the lack of readily available training data. Often they matter less for HEP tasks than in computer science.

2 Exploiting the structure

The most effective way to control the parameter count while keeping expressivity is to exploit the structure of the data. That means if the data to be processed has certain known features or symmetries, the neural network should be adapted to that. The MLP that has been covered so far does not have this capability. It always learns the most generic approximate function based on a set of inputs. While it helps, for example, to sort the reconstructed particles in an event and their variables (or features) before they are fed to the MLP always in the same way, the MLP itself has no structure that in itself accounts for the fact that each particle is a representation of the same physical concept, but possibly with different features. As a universal function approximator, the MLP can learn this connection, however this may require more free parameters and more training (and in turn more computing resources). In some cases, the amount of free parameters that would be needed becomes strictly prohibitive: for example on images. A typical camera e.g. in a smart phone has 10 to 50 megapixels. If each of these pixels would be taken as input to an MLP, and assuming 10 nodes in the first layer, the first weight matrix would contain 100 to 500 Million free parameters. In addition to the large parameter count, the compression to only 10 nodes in the first layer would not provide enough expressivity to perform even simple object identification tasks with such a network. Moreover, if the image was shifted by only one pixel, it would represent a completely unknown input to the neural network even though it shows the same object.

2.1 Convolutional neural networks

A solution to the problem induced by the shifts (and as discussed later for the large parameter count) is to implement filters that can slide over the image. Taking as example the task of identifying an image to show either a cat or a dog, such a cat filter could search for a cat in the image irrespective of the cat's position in the frame. More concretely, such a filter would compare the observed pixel configuration in each part of the image with patterns that could resemble the cat.

Convolutional neural networks (CNNs) [10, 11] implement such filters while keeping the amount of free parameters manageable by (a) learning filters from examples and (b) abstraction. Learning the pixel pattern to look for can be accomplished by a neural network. However, this neural network only considers a fraction of the pixels as inputs: those within the frame the filter is applied to, also referred to as the kernel size. Within such frame, the kernel itself resembles a dense neural network layer, where each pixel corresponds to one input node. This very same layer is applied every time the kernel is shifted. The output of this procedure produces a new image, where each pixel contains the output of the shifted kernel.

2.1.1 Learning filters

In the following we will construct a convolutional kernel step by step, first for a black and white image with one filter:

$$y_j = \theta\left(\sum_i^{N_k} \omega_i \, x_{I(j,i)} - b\right) \tag{19}$$

Here, the value of output pixel y_j is determined by the weights ω (for the moment each ω is strictly a scalar quantity) applied to the pixel values x of the neighbouring pixels. An index mapping function $I(j,i)$ determines the global index of a pixel given a global index of the output pixel j and a relative index within the kernel i. Such a kernel, acting on an input image without bias or activation, is also depicted in Figure 6.

Even though this is a very simple kernel, the main features of a CNN layer already become visible: ω are strictly relative, and each ω is applied and trained multiple times in the same image. This can be extended easily to multiple output nodes (in total N_F), also referred to as output channels:

$$y_{j\alpha} = \theta\left(\sum_i^{N_k} \omega_{i\alpha} \, x_{I(j,i)} + b_\alpha\right), \tag{20}$$

where α determines the output channel. The situation changes slightly when multiple input nodes (also

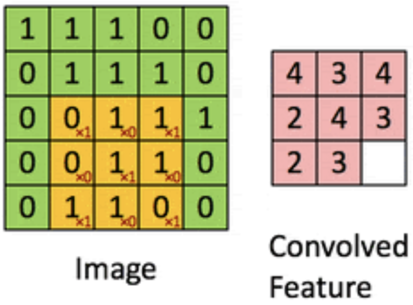


Fig. 6: Illustration of a single filter CNN kernel without bias or activation acting on an input image. The kernel is visualised as red numbers and its size is indicated by the yellow area.

referred to as input channels) are allowed. Then, additionally the input channel (β) has to be accounted for:

$$y_{j\alpha} = \theta\left(\sum_{\beta}^{N_C}\sum_{i}^{N_k}\omega_{i\alpha\beta}\, x_{I(j,i)\beta} + b_\alpha\right). \tag{21}$$

Here N_C is the number of input channels. This equation can be rewritten in terms of matrix multiplications (by – as an exception – introducing vector arrows here):

$$\vec{y}_j = \theta\left(\sum_{i}^{N_k}\omega_i\, \vec{x}_{I(j,i)} + \vec{b}\right), \tag{22}$$

where each ω_i is now a matrix of dimension $N_F \times N_C$ and $\vec{x}_{I(j,i)}$ a vector of dimension N_C. If we unroll the sum further, making ω a matrix with $N_F \times N_k \cdot N_C$ rows and $\vec{x}_{I(j)}$ a vector of $N_k \cdot N_C$ entries, determined by $I(j)$, we arrive back at an expression similar to the dense layer from Equation 1:

$$\vec{y}_j = \theta\left(\omega\, \vec{x}_{I(j)} + \vec{b}\right). \tag{23}$$

However, the relative nature of the convolutional kernel is still expressed by the pixel index j and the fact that also $\vec{x}_{I(j)}$ is built by selecting pixels relative to j through $I(j)$.

In this form, the connection to a discrete convolution, defined by:

$$(f * g)[n] = \sum_{m=-\infty}^{+\infty} f[m]g[n-m] \tag{24}$$

is not as clearly visible. However, one can relate the expression above to a convolution. The main part is to change the perspective of the term $\omega_i\, \vec{x}_{I(j,i)}$ in Equation 22 to refer to all N_p pixels rather than just the kernel:

$$y_j = \sum_{m=1}^{N_p}\omega(j,m)x_m, \tag{25}$$

where $\omega(j,m)$ returns the weight for global pixel index m within the kernel, if m is within the frame around j and zero otherwise. This can also be expressed to be a function of the difference between j and

m, and therefore it can be related to a convolution as follows[1]:

$$y_j = \sum_{m=1}^{N_p} x_m \, \omega(j - m). \tag{26}$$

As the CNN layer is indeed equivalent to a convolution, the property of *translation equivariance* also follows directly from it since translation and convolution operators commute, meaning that applying a pixel (or coordinate) shift and then the CNN layer returns the same output as applying the same CNN layer and then shifting the image. This should not be confused with translation invariance, where the output of the operation does not change with the translation. The latter property can apply, for example, to the whole neural network that classifies cat versus dog pictures (that will likely contain also CNN layers).

Special care needs to be taken at the edges of the images, that we so far neglected. As also indicated in Figure 6, the CNN layer would reduce the image size as the kernel cannot extend beyond the image boundaries. In some cases, however, the output should have the same size as the input (for example when denoising an image). In this case, one can apply a simple padding to the edges, where pixels with a fixed value (usually zero) are added such that the application of the CNN layer does not reduce the image size.

2.1.2 Abstraction and pooling

Even if a filter does not need to process the whole image, a kernel that is able to identify e.g. a cat would still need to cover a large area of the image, leading to a large kernel size, and that resulting in a large number of free parameters: $N_C \cdot N_F \cdot N_k$. In CNNs, this is kept under control through abstraction and pooling. The kernel sizes at each CNN layer are kept small, e.g. the first layer would only identify edges in an image, changes from one colour to the other, in very localised areas. Before the next layer, one would apply a pooling operation on this resulting image. A pooling operation summarises the information of neighbouring pixel groups by taking the mean or maximum value (the latter is more commonly used) for each feature over that pixel group. This way, the information contained in the image is compressed more and more. At the same time, this introduces abstraction into the CNN, since now the next layer would combine edges, represented by hidden vectors per pixel, into more complex objects, such as a cat's eye. Then subsequent layers could identify two cat eyes (accompanied by the corresponding mouth and nose) as a cat face etc.

Without introducing a large amount of free parameters in the early layers, this architecture gives rise to a large *receptive field* of the deeper layers of the network, illustrated in Figure 7, which finally allows a kernel in a deeper layer to "see" a whole object (in terms of the combination of all its parts) at once, and therefore identify it.

The architecture of a classification CNN is shown in Figure 8. Even though it is an early example, it shows all the typical features discussed above: it contains convolutions over the input image or hidden

[1] N.B.: there is a subtle flip in the matrix ω in this step that can be fully absorbed by the fact that the parameters of ω are free learnable parameters. Technically speaking the convolution operation in neural networks is actually a cross-correlation operation.

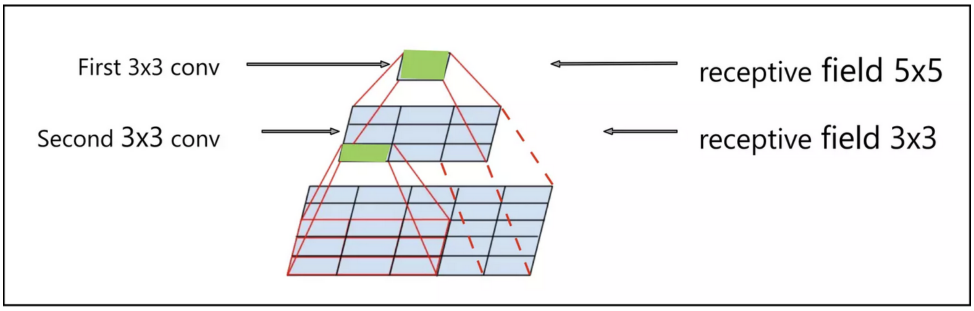


Fig. 7: The receptive field of a hidden layer in a CNN.

representations intersected with pooling operations, building a more abstract representation of the image. Since the final goal of the network is to identify hand-written numbers, the output of the last pooling operation is transformed into a simple vector (flattened) and passed to a dense neural network, finally classifying the number to be between 0 and 9.

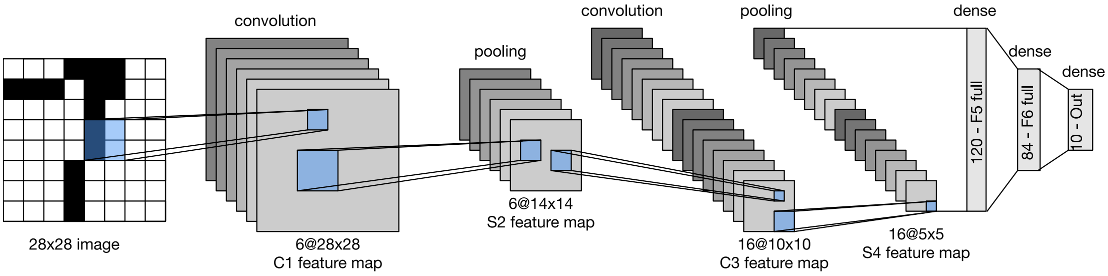


Fig. 8: “LeNet” [11].

To summarise, CNNs rely strongly on CNN layers. Each CNN layer is translation equivariant and therefore exploits the structure of image data. Moreover, the same weights are applied and trained multiple times within an image. This also means that the effective data set each weight can be trained with is up to multiple orders of magnitude larger than the number of samples contained in the data set, reducing the risk of overtraining significantly.

Examples of a direct application of a CNN in the HEP context can be found in Ref. [12] and many others. In particular the identification of the origin of jets is here a topic that has received a lot of attention at about the same time advanced DNN structures became used more extensively in HEP. In Reference [12], the energy deposits of the jet in the calorimeter are interpreted as an image. Since many current calorimeters can be segmented regularly in η and ϕ, a regular pixel grid can be built, centered around the jet direction, which is then fed to a classification CNN to identify the jet origin. Here, jets that stem from top quarks and jets that stem from lower-mass QCD interactions are being distinguished.

Figure 9 shows these jet images. In the preprocessed case, there is a clear distinction visible for the

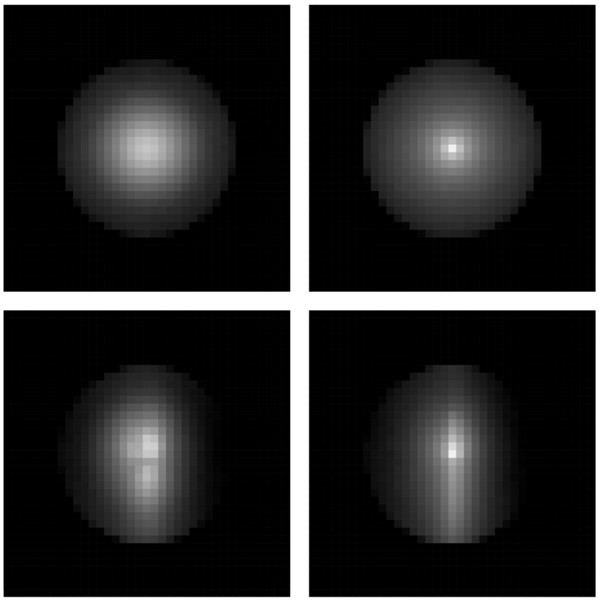


Fig. 9: The average of 100k jet images. The gray-scale intensity corresponds to the total transverse momentum in each pixel. Upper: no preprocessing besides centering. Lower: with processing (rotations and flips) ensuring the maximum intensity is in the upper right quadrant. Left: top-quark jets. Right: background jets. Figure taken from Ref. [12].

top-quark and the background jets that can be exploited by the CNN, improving the performance with respect to the case, where only high-level variables constructed by humans are used. A different way of using the structure of the data to identify jets - in this case stemming from b or c quarks - is incorporated into DeepJet [13]. This jet identification algorithm uses the full set of jet constituents, individually reconstructed first using all subdetectors. Each of these particle candidates is then processed by a per-particle neural network, which can be interpreted as a CNN layer with a kernel that operates on only one pixel (= particle) at the time, before the output is flattened and fed to a feed-forward DNN [2]. Exploiting the structure of the data in this way led to a tremendous improvement of the physics performance, in particular at high transverse momenta, compared to previous MLP approaches that needed to restrict themselves to selecting a few most significant jet constituents as they did not directly mirror the data structure.

2.2 Attention and transformers

In particular in HEP, most of our data does not come in the form of a regular grid - or is easily transformable to one. However, a regular grid, and some sort of translational symmetry on that grid, is needed to use CNNs. A prominent example from outside of HEP for data that does not follow a regular grid structure, but is also not fully unordered, is language. Sentences are sequences, where the order of the words determines their meaning, together with their individual meaning. So to match, for example, an image of a bear looking at a trout in a river to the sentence “The little bear saw the fine fat trout in the brook.”, a CNN for the image, and a mechanism to interpret the written sentence is needed. Also here,

[2] N.B: Also a recurrent neural network layer is used in-between the per-particle network and the feed-forward part, however the per-particle network was the main enabler

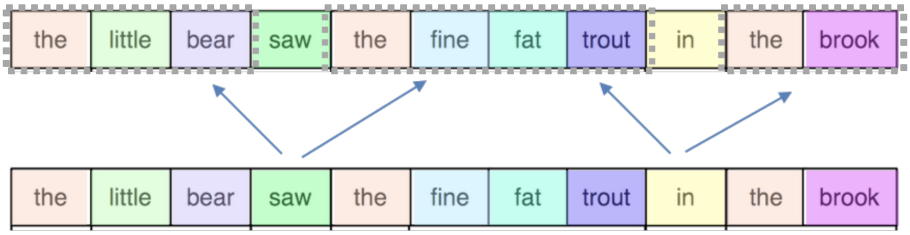


Fig. 10: Example relations between words in a sentence. The embeddings are symbolised by different colours, and the relations by arrows.

the structure of the data is the key. For sentences, usually the following applies:

- they are made of words, not pixel-values
- they do not have a fixed length
- the relations between the words matter
- there is a direction

In the following, we will go through these structural properties and how they are addressed in language-processing models starting from the CNNs to finally arrive at the *attention* mechanism, which is at the heart of the *transformer* architecture, the basis for powerful models such as chatGPT.[3]

In order to transform words into numbers that can be processed by a DNN, an *embedding* is performed, which is simply a translation of words (or parts of words) through a dictionary to vectors of real numbers. These embeddings can simply be learned by a shallow neural network that translates an index in the vocabulary (a number assigned to each word) to a vector in a higher dimensional space that is normalised to have a length of one to provide reasonably normalised input to the following part of the neural network. In addition to a word embedding, also a position embedding can be performed that would be added to the input feature vector of each word. More details on position embeddings can be found e.g. in Ref [14], but are not discussed in detail here since they are less relevant for the HEP context.

What is similar between an image and a sentence is that often words that relate to each other are actually close together in the sequence. For the issue of a fixed length one could - in principle - add zeros up to a maximum fixed length. This would allow to use the concepts of CNNs also for sentences: filters, abstraction, and summary building (pooling). However, while this may work for the short example here, it will not work for very long sentences, in particular for languages where words that relate to each other do not have to be close to each other in the sentence. However, it is possible to extend this idea in the sense that if it is not the nearest neighbours that are necessarily relevant, then a way to determine the relevance of the neighbours could be by the relation of one word to the other words in the sentence. This is illustrated in Figure 10.

So similar to a CNN layer, where the value of output pixel j is determined by the input values x of pixels in its neighbourhood (see e.g. Eq.22), one can define an operation here, that takes into account

[3] This is a very non-historic approach and often taught differently in the literature.

the relation a between the words to define an output value:

$$\hat{y}_j = \sum_i^N a(i,j) x_i. \tag{27}$$

This relation should depend on the words themselves, and possibly the position of the word in the sentence, in other words the input features of each word. Therefore, we can substitute $a(i,j)$ with $a(x_i, x_j)$. In addition, this *attention* should be directed as one for example the word "in" may need to relate to "trout" differently than "trout" would relate to "in". This can be accomplished by defining a function $k(x)$ that creates a *key* from the word features and a function $q(x)$ that creates a *query* from the word features. Making these functions learnable (e.g. by introducing a simple weight matrix), the attention can be expressed as $a(k(x_i), q(x_j))$. In addition, the attention should be scalar, while the function k and q are more expressive if they are vector-valued. This leads to the final definition of the attention (often referred to as dot-product attention):

$$a(i,j) = \underset{i}{\sigma} \left(k(x_i) \cdot q(x_j) / \sqrt{d_k} \right), \tag{28}$$

where $k(x_i) \cdot q(x_j)$ is the scalar product between the vectors $k(x_i)$ and $q(x_j)$, also referred to as *alignment*. The function $\underset{i}{\sigma}$ is the softmax function:

$$\underset{i}{\sigma}(\xi) = \frac{\exp(\xi_i)}{\sum_j \exp(\xi_j)}, \tag{29}$$

ensuring that the sum of the attention weights is one. The normalisation term $\sqrt{d}$ ensures that the variance of the scalar product remains roughly at one (with d being the dimensionality of the input vectors, leading to a variance of the scalar product of d for normal distributed vectors). The full expression then reads as:

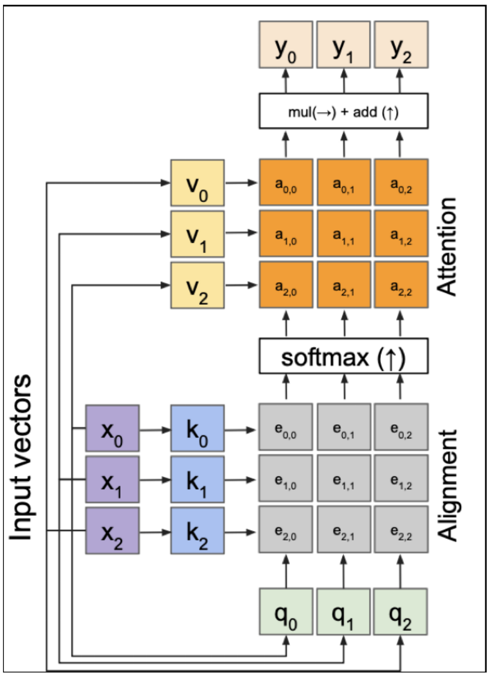
$$a(i,j) = \underset{i}{\sigma} \left(\sum_\alpha^{d_k} \left(\sum_\beta^{N_C} \omega_{\alpha\beta} x_{i\beta} \right) \left(\sum_\gamma^{N_C} \tilde{\omega}_{\alpha\gamma} x_{j\gamma} \right) / \sqrt{d_k} \right), \tag{30}$$

where ω and $\tilde{\omega}$ are the weight matrices for k and q respectively and N_C is the number of input features per word (analog to the number of input channels for a CNN layer). To add more expressivity, also the features of each word x are transformed before they enter the attention-weighted aggregation, using another weight matrix, defining a full attention operation [15]:

$$y_{j\alpha} := \sum_i^N a(i,j) \sum_\beta^{N_C} \hat{\omega}_{\alpha\beta} x_{i\beta} = \sum_i^N a(i,j)\, v_{j\alpha}. \tag{31}$$

The last term is often referred to as the "value" transformation, making the attention aggregation work on *key* (k), *query* (q), and *value* (v) inputs. The whole mechanism here expressed as equations is also illustrated in Figure 11. Since key, query and value are all derived from the word features themselves, this operation is also referred to as self attention[4].

[4] This has also historical reasons.



y_0 y_1 y_2
mul(→) + add (↑)
v_0 v_1 v_2
$a_{0,0}$ $a_{0,1}$ $a_{0,2}$
$a_{1,0}$ $a_{1,1}$ $a_{1,2}$
$a_{2,0}$ $a_{2,1}$ $a_{2,2}$
Attention
softmax (↑)
Input vectors
x_0 x_1 x_2
k_0 k_1 k_2
$e_{0,0}$ $e_{0,1}$ $e_{0,2}$
$e_{1,0}$ $e_{1,1}$ $e_{1,2}$
$e_{2,0}$ $e_{2,1}$ $e_{2,2}$
Alignment
q_0 q_1 q_2

Fig. 11: Illustration of the self-attention mechanism.

Similar to defining multiple filters in a CNN layer, also here multiple *attention heads* can be defined, each attending to a different set of words, by repeating the above operation with independently learnable weight matrices. The respective outputs of the heads are then simply concatenated, meaning that the individual vectors, e.g. N vectors of dimension M, are then combined to a single vector with $N \cdot M$ entries. We can see how this mechanism could be immediately useful, e.g. when processing a jet with multiple constituents that do not follow a regular grid pattern, or other inputs in HEP. However, the most prominent application of the attention mechanism are transformer models.

A transformer is a neural network architecture that is built for sequence processing and revolves around the application of attention blocks. Due to its current prominence, there are many excellent sources on the internet that explain it very well, which have partially animated graphics, one of those being Ref. [16]. Therefore, the description here is kept a bit shorter and points out in particular those details that can also help improve the neural network performance in other architectures. A general overview of the architecture is given in Figure 12. The transformer model consists of an *encoder* block, that transforms the input to a hidden representation, and an *auto-regressive decoder* block that serves two purposes: it produces an output probability for the next word to be produced based on the encoded input text (e.g. when translating text from one language to another) and reads back its own output in an auto-regressive manner to produce the following words, conditioned on the fact that it has already produced the first (or a set of) words. It is important to note that in the connection between encoder and decoder, the keys and values of the encoder are fed to the second multi-head attention block of the decoder, such that the decoder can query different parts of the encoded input depending on the words that have already been generated.

Most building blocks of the transformer architecture in Figure 12 were already discussed. However, the "Add & Norm" blocks require additional explanation, in addition to the *skip connection* that

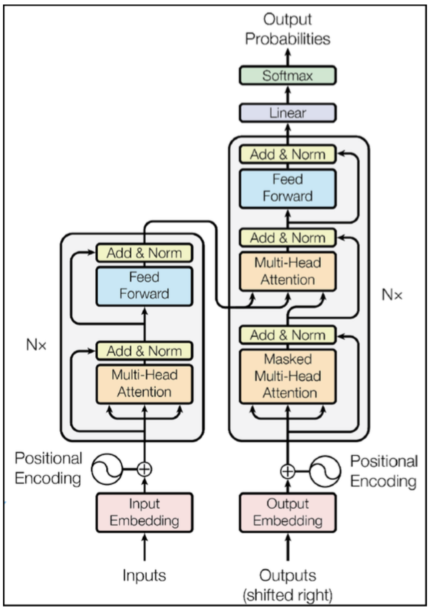


Fig. 12: Illustration of the original transformer architecture, taken from Ref. [15]. The “Nx” stands for the possibility to repeat the respective block multiple times.

connects the inputs and the outputs of each multi-head attention layer. A skip connection is a powerful tool when designing neural networks. It allows the information to flow through the operation that is skipped, but also to skip it. If, as in the case of the transformer, the output of the operation (here the multi-head attention block), is added to the original features, the operation only adds corrections to the original features, which is often easier to learn in practice. Moreover, it aids the training as there is always a direct connection of the gradients from the loss value to each of the learnable weight matrices in the backward pass. In consequence, it is much less likely that gradients vanish or explode. In addition, a *layer normalisation* is applied in the transformer model, which normalises each feature vector to have a magnitude of one, which can also help the neural network to learn faster and more reliably.

For sequence processing, the sequence is then fed to the transformer model, concluded with an *end token*. An end token is part of the vocabulary that marks the end of a sequence. Once the information is processed, the decoder part starts creating outputs. Here, the most probable word is chosen and passed back to the decoder into the auto-regressive input. Then the decodes produces the remaining part of the output sequence iteratively until an end token is produced by the decoder, at which point the generation ends.

A rather recent application of a transformer model in a HEP context is targeting jet identification, which remains a challenging task. The developments are heavily aided by readily available data sets to train and evaluate the models on, however I expect less and less relative improvement from newer models as the existing ones are already exploiting the information available to them quite well. The application to jets in HEP is called Particle Transformer [17]. The constituents of the jet are here interpreted as words in a sentence, each with its own features such as momentum and particle type. The final output of the architecture is a prediction of the particle that originated the jet, so a single (vector valued) output encoding the jet class instead of a sequence in the original transformer model. This is achieved by passing exactly one (empty) class token to the decoder part to create the queries of the decoder to the encoded inputs finally leading to the prediction of the jet class. There is second main aspect that differs from the original transformer model: in the particle transformer the alignment matrix (the dot product of key and query pairs) is altered. In addition to fully learned key and query pairs, the physical relations of pairs of constituents in the jet, such as the angles between them, are added before the softmax function is applied. This decreases the mis-identification rate at a fixed efficiency to correctly identify the jet by up to a factor of almost two. Therefore, the particle transformer network is also an example that adding structural information (in particular physics-motivated information) to the neural network can help improve the performance - and with it the final physics reach - significantly.

2.3 Graph neural networks

In addition to image-like data and sequences, data in HEP can also come in other forms. One of them is the form of a graph. Formally, a graph consists of a set of points, also referred to as nodes or vertices, and connections between them, referred to as edges. These edges can either be undirected or directed. A prominent example for an undirected graph is the Facebook friend network, where friend requests have to be accepted by both parties. An example for a directed graph is the Twitter network, where person A can follow person B without B needing to follow A. This also illustrated that in a graph vertices as well as edges can have properties. Therefore a graph is an abstract way to describe “things” and relations between them.

In order to define learnable operations on such a graph a few things need to be considered. A graph does not provide a particular ordering of its content, therefore sequence processing operations cannot be applied directly. There is also no regular grid structure in a graph, so also CNN-like approaches cannot be applied, and finally a graph does not need to be of fixed size so even if it were computationally feasible, a simple MLP is also not the right tool to process graph information.

However, the graph itself provides a structure that can help define operations on it: the simplest operation is to independently update edge or node information, without any information exchange across the graph. In addition, information can be exchanged along edges between nodes, in a directed or undirected way, and possibly influenced by the edge properties. Moreover, edges or nodes can exchange information with a global entity. These paths of information exchange are summarised very well in Ref. [18]. Since the graph nodes and edges have no particular ordering, the concrete implementation of these operations should respect order invariance. This limits their set basically to a sum (or mean), a product, or minimum and maximum operations.

The simplest information exchange in a graph, where nodes are connected to other nodes is the following:

$$y_j = \underset{i\,\epsilon\,N(j)}{\square} x_i, \tag{32}$$

where node j is updated by taking the sum (or mean), product, minimum, or maximum (here symbolised by $\square$) over the set of connected nodes $N(j)$. Another way of expressing this operation is using the *adjacency matrix* A:

$$y_j = \underset{i}{\square} A_{ji} x_i. \tag{33}$$

Here $A_{ij} = 1$ means that two nodes are connected and the $\square$ is applied over all nodes. More generally this *message passing* can be expressed as:

$$y_j = \mathop{\square}_{i \,\epsilon\, N(j)} \Phi(x_j, x_i), \tag{34}$$

where Φ is a generic function or a neural network. In this form, this reminds strongly of the attention mechanism introduced earlier, and indeed an attention layer can be interpreted as a weighted information exchange between nodes, with edge weights given by the attention between the nodes. Even a CNN can be phrased as a graph neural network, where the edge weights are learned and depend statically on the spatial relation of the central node (pixel) to the neighbouring nodes (pixels) - with the exception that here there is a specific ordering that can be exploiting directly. This illustrated that the graph formalism is very powerful and many conclusions drawn here apply directly to other neural network types.

In the HEP context, however, often there is not information to build a complete graph. In particular the edges are often missing. For example, the constituents in the jet or generally reconstructed final state particles in an event a-priori do not have connections between them. The same is true for hits in a detector, where a particle induced a signal above threshold, which have a position and other properties but no connections that would define a graph. This data mostly resembles a *point cloud*, a set of unordered points with properties.

So in these cases, a graph needs to be defined before operations can be applied on this point cloud. A simple solution would be to connect each point with all other points. This can only work in very simple cases, where the number of input points, N, is low, since it makes the number of connections grow by N^2, which makes this approach quickly resource-prohibited. Another approach could be to connect all points within a certain radius in the physical space. But also here, it is unclear if the connections that are defined are actually optimal for solving the problem. It can easily lead to too many connections or to few connections that can result on limited information exchange across the graph. It has been shown even formally that even a deeper graph with more iterations of information exchange cannot mitigate the issue of information-passing bottlenecks introduced by non-optimal connections [19] (and earlier references therein).

In the following these notes will focus on two examples of neural networks with learnable connections that optimise the graph topology during training. One of them is a the dynamic graph convolutional neural network (DGCNN) [20]. Input to the DGCNN architecture are a set of points making up the point cloud that may contain additional properties. Each of these vectors of point features is first transformed by a point-wise MLP, building a higher dimensional representation of each point. Based on this input there are two key features in the DGCNN: first, k nearest neighbours (in the original paper around 64) are selected using the euclidean distance between the points in full feature space; then the *EdgeConv* operation is applied:

$$y_j = \mathop{\square}_{i \,\epsilon\, N(j)} \Phi(x_i, x_j - x_i), \tag{35}$$

where Φ is an MLP and for $\square$ the maximum is used. The key element here is that the operation processes the relations between the points, in other words the edge properties, where the edge properties are $e(i, j) = x_i - x_j$. Depending on the task, either the node properties are then aggregated into one global descriptor (a global vector built by maximum pooling over all nodes) and this global vector is used as

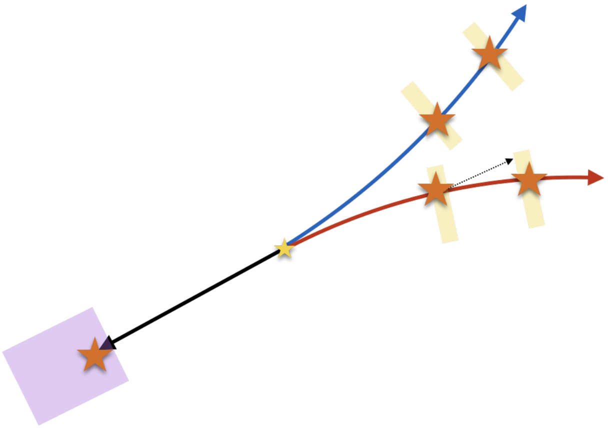


Fig. 13: Sketch of particles produced in a central collision passing different detector elements. Different particles are indicated by red, blue, and black arrows. A calorimeter is illustrated by the purple box, and tracking detector layers by yellow boxes. The interaction with the detector elements is symbolised by orange stars and the primary event with a yellow star.

input to a feed-forward DNN to perform a classification task of the entire point cloud, or the global vector is added to each node feature again, e.g. to perform a *semantic segmentation* task, where each point is assigned to be part of a certain class of object (e.g. of a car or a pedestrian).

This architecture is very powerful and has been adopted for many applications in HEP, such as ParticleNet [21]. ParticleNet is equivalent to the DGCNN architecture modulo small modifications regarding the choice of a few hyper parameters applied to the identification of jets[5]. In comparison to, for example, the image-based jet identification algorithms, using the DGCNN architecture improved the performance drastically, owing to the fact that it captures the structure of the jet data, containing individual constituents of the jet rather than only images, better. Moreover, the EdgeConv operation, targeting the relation between the constituents in the jet directly, contributes to the improved performance, and is another example that exploiting the structure of the data almost always comes with benefits.

3 Examples for advanced applications in HEP

While many of the techniques discussed here target classification, e.g. the separation of a signal from a background process with feed-forward neural networks or the classification of jets, there are also other applications of machine learning in HEP. One of them is closely related to the classification task: the regression task, e.g. to regress a correction to the measured jet momentum. These are conceptually straight forward extensions of classification neural networks and the same considerations apply. However, there is another class of applications that are conceptually very different: the reconstruction of multiple objects from detector signals, also illustrated in Figure 13. On the one hand, these tasks require only a single output, or a fixed size output; and on the other hand, they pose additional constraints on the resources that are not as severe in other cases. Both will be discussed in the following.

A typical modern multi-purpose HEP detector such as CMS or ATLAS consists of multiple subde-

[5] the focus on jet identification in these lectures has two reasons: one is that a lot of advanced machine learning techniques have found their way into this area, and the other one is to keep consistency between the applications that allows to compare them given the limited scope of these lectures.

tectors, each equipped with a large number of possible read-out channels fed by active detector elements that can detect particles traversing them or being stopped. In total, such a detector can easily reach $(O)(10^8)$ sensors. With an occupancy in the range of per-cent for a typical event, this means $(O)(10^6)$ of such active sensors, also referred to as hits, need to be processed.

While the detector data lends itself to being interpreted as a point cloud, and graph neural networks are very powerful in extracting information from that point cloud, the large amount of input points needs to be considered in the choice of the neural network architecture and can easily pose prohibitive constraints on it. For example if a self-attention-based layer such as in a transformer would be chosen, N^2 attention weights would need to be calculated. With $(O)(10^6)$ points this is not feasible. Here, more local algorithms provide two advantages: they only consider a smaller amount of possible connections and are therefore more likely to be computationally feasible; and they also have a better chance of being robust when confronted with unknown (physics) data, since they are strictly local. Locality is an important feature in such algorithms since the reconstruction of one particle should not be affected by the reconstruction of another particle at the other end of the detector[6]. For example DGCNN could provide such a locality constraint as it only processes edges of nearest neighbours. As a reminder, the main operations in DGCNN are building K nearest neighbours in the feature space of dimension F, and processing the edge information between the neighbours, also of dimension F. That means, with N points as input, one iteration of EdgeConv subjects an MLP to $N \times K \times F$ inputs, where the MLP has multiple layers with roughly F^2 free parameter. For typical values that would amount to more than $100\,000 \times 64 \times 64^2$ parameters to evaluate. This poses an issue that is overcome by GravNet [22]. As a first step, a GravNet layer learns coordinates of small dimensionality S from the input features, and additionally a feature vector to facilitate the information exchange. The first advantage in terms of resources is, that only the low-dimensional coordinates are used to calculate the k nearest neighbours, bringing a factor of 10 improvement in terms of resource usage over DGCNN. The second advantage is that instead of processing the edge information, in GravNet features are aggregated by a weighted sum - without MLP operations on the edges:

$$y_j = \underset{i\,\epsilon\,N(j)}{\square} \exp(-d_S(i,j)^2)D(x_j), \tag{36}$$

where $d_S(i,j)$ is the euclidean distance between node i and j in coordinate space S. This implements a geometric attention mechanism with D being a dense layer creating the value to be passed to the attention operation. The Gaussian weighting ensures that small distances correspond to large attention. In this way, points that should exchange a lot of information are pulled together in the coordinate space. In turn this means selecting the k nearest neighbours in that space also builds an optimal graph for the task at hand, since finally the loss function will determine what the best path of information exchange is. Moreover, this phrasing avoids processing edge information by any learnable operation directly, leading to a reduction of resource needs with respect to EdgeConv of a factor $k \approx 64$[7]. The full sequence of operations is also illustrated in Figure 14.

The physics performance of this information exchange is comparable or at times even better than

[6] N.B: Locality is also important for calibration of said algorithms, their bias they receive from the training sample, and their interpretability.

[7] In practice this can become even larger due to the way algorithms are implemented

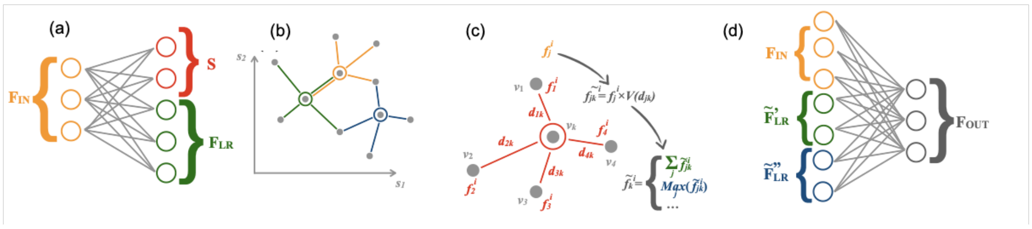


Fig. 14: Illustration of the steps in the GravNet layer. a) The input features are projected to coordinate features, S, and features to be exchanged between the points (F_{LR}. b) Nearest neighbours are connected in the space spanned by the coordinates S. c) The neighbour features are aggregated weighted by the distance in space S, taking the mean and the maximum of the weighted neighbour features. d) The aggregated information and the original inputs are fed to a dense neural network layer to build the final output of the GravNet layer.

the performance of DGCNN, for example when assigning hits in a highly granular calorimeter to one of two particle showers that stem from charged pions, or when assigning hits to different classes of objects in events, as in a study by the ATLAS collaboration [23]. This is noteworthy because it enables full event reconstruction with machine learning from detector hits in terms of computing resources, as well as that it showcases the strength of attention-based architectures together with the possibilities arising from learning the graph topology that facilitates the information exchange directly.

As just discussed network architectures exist for the reconstruction of multiple particles from detector hits, however these networks need to be trained, and individual objects need to be separated. This task is usually referred to as *instance segmentation* in computer science (not to be confused with *semantic segmentation*, where the aim is to determine the class of the object a certain point belongs to).

In HEP most computer vision techniques to perform such an instance segmentation do not apply directly [24]. This can be achieved, however, with object condensation [24], a technique that trains the neural network to convert a complex problem like separating the hits of different particles in the detector, that can overlap partially and have complex shapes, to a simple problem where the hits are reorganised in a learnable space and build clear ideally round clusters. In addition, also properties can be learned for each object such as the particle type or its momentum. Without going into too many technical details, the central point is to define a confidence measure β for each point and train it such that at least one point per object has a large β value and all noise points have low β values. Then, potentials are defined that create an attractive force for points belonging to the same object in the learnable *clustering space*, and a repulsive force between points that belong to different objects. To pair-wise loss calculation that would lead to $\mathcal{O}(N^2)$ relations to be calculated, only the highest β points per object are considered to act on all other points, but the remaining points do not directly interact with each other. These attractive and repulsive potentials created by each object are scaled with the β value of these highest points of the corresponding objects. They also scale with the β value of point that is either attracted of repelled, such that the gradient that is calculated to train the neural network creates a force scaling with β. In consequence, the highest β points will be in the center of the clusters in clustering space. Moreover, the network can also be trained to predict the object properties. They are predicted for each point belonging to the that object, and the corresponding loss is also scaled with β. As a result, the central highest β points per cluster also carry the best property estimate for the object. They are referred to as *condensation* *points*. In a last step, the objects can be collected in clustering space by starting with the highest β point, assigning points around it (within a distance threshold) to that object and removing them from further processing, and moving on to the next highest β point until another threshold is reached. This simple last clustering step then ensures no object duplicates. This final step is illustrated in Figure 15. In particular visible is that in the clustering space, the cars are well separated from the background pixels and from each other. Even though the algorithms are not made for the task shown here to identify and segment individual cars in the images, they perform reasonably well, and are comparable to state-of-the-art computer vision approaches.

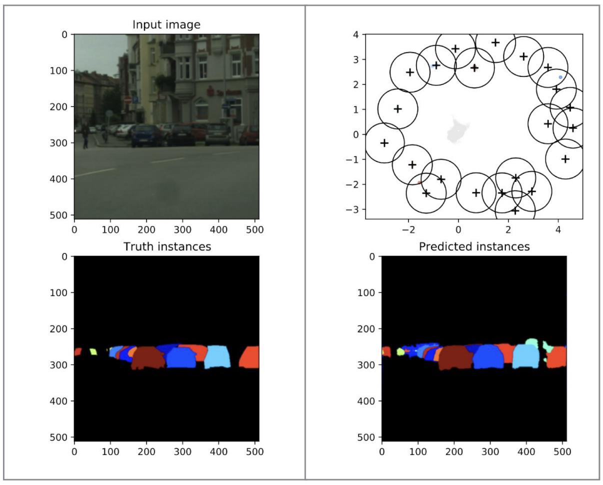


Fig. 15: Application of object condensation to a computer vision problem, identifying cars in an image. Top left: the original image, top right: the clustering space including the circles assigning points to each of the condensation points. Here, the background pixels are coloured grey. Bottom left: the truth instances, assigning each pixel to a particular car or the background. Bottom right: cars segmented by object condensation.

In the HEP context, this technique has also proven to be very powerful in context of calorimeter reconstruction (e.g. Ref. [25] and others), but also for track reconstruction [26]. Moreover, paired with neural networks like GravNet it allows creating a rather detector agnostic reconstruction that could be applied (after retraining) to very different types of detectors or problems.

Besides architectures that exploit the structure of the problem, there is a plethora of applications of such architectures for many purposes beyond classification or object detection. Of particular interest for physics are anomaly detection, often relying on auto-encoders, or even generative networks that can be used to generate images, but also as a fast simulation of the interaction of particles with matter, targeting future computing challenges. The underlying principles from a machine-learning perspective are described in a concise form in Ref. [1] and the references therein. Their applications to high energy physics are listed in the living review of ML for HEP [27].

4 Summary

Machine learning has left a tremendous impact and has enhanced the physics reach in high energy physics in the last years significantly. In most cases, this success relies on the exploitation of the structure of the problem (the physics) while giving sufficient freedom to the algorithms to learn also subtleties of the data that are inaccessible for classic algorithms. This lecture aimed particularly at providing a tool set to mirror the physics structure in the network architecture, covering MLPs, up to attention based algorithms, and also provided a glimpse of advanced object detection in physics detectors.

References

[1] F. Fleuret, The little book of deep learning, https://fleuret.org/francois/lbdl.html.

[2] S. Dubey *et al.*, Activation functions in deep learning: A comprehensive survey and benchmark, doi:10.1016/j.neucom.2022.06.111.

[3] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics.

[4] K. He, *et al.*, Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV).

[5] I. Goodfellow, *et al.*, Deep Learning, MIT Press (2016).

[6] L. Ljang, A Visual Explanation of Gradient Descent Methods, https://towardsdatascience.com (2020).

[7] A. G. Baydin, *et al.*, Automatic differentiation in machine learning: a survey, arXiv:1502.05767 (2015).

[8] Stanford CS231n: Deep Learning for Computer Vision (MIT Licence), https://cs231n.github.io/neural-networks-3.

[9] J. Jordan, Setting the learning rate of your neural network (2018), https://www.jeremyjordan.me/nn-learning-rate/.

[10] K. Fukushima, Neural network model for a mechanism of pattern recognition unaffected by shift in position - Neocognitron. (1979), Trans. IECE

[11] Y. LeCun, *et al.*, Gradient-based learning applied to document recognition (1998), Proceedings of the IEEE 86, doi: 10.1109/5.726791

[12] S. Macaluso, D. Shih, D. Pulling out all the tops with computer vision and deep learning, JHEP (2018) doi: 10.1007/JHEP10(2018)121

[13] E. Bols, *et al.*, Jet flavour classification using DeepJet, JINST 15 (2020) doi: 10.1088/1748-0221/15/12/P12012

[14] K. Erdem, Understanding Positional Encoding in Transformers (2021), towardsdatascience.com

[15] A. Vaswani, *et al.*, Attention Is All You Need, NeurIPS proceedings 2017

[16] Jay Alammar, The Illustrated Transformer, jalammar.github.io

[17] H. Qu, *et al.*, Particle Transformer for Jet Tagging, Proceedings of ICML2022

[18] P. Battaglia, *et al.*, Relational inductive biases, deep learning, and graph networks, arXiv:1806.01261

[19] F. DiGiovanni, *et al.*, On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology, Proceedings of ICML2023

[20] Y. Wang, *et al.*, Dynamic Graph CNN for Learning on Point Clouds, ACM Transactions on Graphics (2019) arXiv:1801.07829

[21] H. Qu, L. Gouskos, ParticleNet: Jet Tagging via Particle Clouds, Phys. Rev. D 101 (2020) doi: 10.1103/PhysRevD.101.056019

[22] S. Qasim, *et al.*, Learning representations of irregular particle-detector geometry with distance-weighted graph networks, EPJC (2019) doi: 10.1140/epjc/s10052-019-7113-9

[23] ATLAS Collaboration, Physics Object Localization with Point Cloud Segmentation Networks (2021), ATL-PHYS-PUB-2021-002

[24] J. Kieseler, Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph, and image data, EPJC (2020) doi: 10.1140/epjc/s10052-020-08461-2

[25] S. Qasim, End-to-end multi-particle reconstruction in high occupancy imaging calorimeters with graph neural networks EPJC (2022) doi: 10.1140/epjc/s10052-022-10665-7

[26] K. Lieret, *et al.*, High Pileup Particle Tracking with Object Condensation (2023), doi: arXiv:2312.03823

[27] HEP ML Community, A Living Review of Machine Learning for Particle Physics https://github.com/iml-wg/HEPML-LivingReview

Heavy-ion physics

Korinna Zapp[a]
[a]Lund University, Lund, Sweden

Collisions of heavy ions at collider energies provide us with a unique opportunity to study strongly interacting matter at extreme temperatures and densities in the laboratory. Under these conditions quarks and gluons become deconfined to form a new state of matter, the quark-gluon plasma. Heavy ion physics has seen three major discoveries in the last 30 years. The first is that the QGP is the least dissipative material known and behaves like an almost perfect liquid. The second is that jets which are the manifestations of highly energetic quarks and gluons are strongly suppressed and modified compared to proton–proton collisions. This so-called jet quenching can be understood as the partial equilibration of a far-from-equilibrium system in a thermal QGP. The third main discovery is that particles with low transverse momentum produced in small collision systems like high multiplicity proton–proton and proton–ion collisions show many features that were believed to be signs for QGP formation. On the other hand, no jet quenching has been observed so far in small collision systems. These lectures are meant to give an overview over all relevant aspects of heavy ion physics at a phenomenological level.

1 Introduction

A large number of particles is produced in collisions of heavy atomic nuclei at present day colliders. In Pb+Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 2.76\,\mathrm{TeV}$ in CERN's Large Hadron Collider LHC, for instance, up to 1600 primary charged particles are produced per unit rapidity in the central rapidity region[1]. These particles originate from a system of very high energy density, and following an argument by Bjorken [1] the initial

This chapter should be cited as: Heavy-ion physics, Korinna Zapp, DOI: 10.23730/CYRSP-2025-009.215,
in: Proceedings of the 2023 European School of High-Energy Physics,
CERN Yellow Reports: School Proceedings, CERN-2025-009, DOI: 10.23730/CYRSP-2025-009, p.215.

[1] For comparison, the number of primary charged particles per unit rapidity in typical proton-proton collisions at the same $\sqrt{s}$ is only around 4–5.

energy density shortly after the collision of the nuclei can be estimated from the radius of the nuclei R, an initial proper time $\tau_0 = \mathcal{O}(1\,\text{fm/c})$ and the measured transverse energy at mid-rapidity as

$$\epsilon_0 \simeq \frac{1}{\pi R^2 \tau_0} \left. \frac{\mathrm{d}\,E_\perp}{\mathrm{d}\,\eta} \right|_{\eta=0} \simeq 25\,\text{GeV/fm}^3 \tag{1}$$

for the scenario described above [2]. This corresponds to roughly 25 times the density inside a proton. In a strongly interacting system of such high density there has to be re-scattering in the final state, and scattering drives a system towards thermal equilibrium. *It is one of the central questions in heavy ion physics how and to what extent the system produced in heavy ion collisions thermalises.*

The colliders with heavy ion programs at high beam energies are currently the Relativistic Heavy Ion Collider RHIC at BNL on Long Island and the Large Hadron Collider LHC at CERN in Geneva. RHIC accelerates a variety of different ions to centre-of-mass energies of up to $\sqrt{s_{\text{NN}}} = 200\,\text{GeV}$, i.e. $200\,\text{GeV}$ per nucleon pair. The largest data sets are for Au+Au collisions. The heavy ion program at the LHC has so far focused on Pb+Pb collisions at centre-of-mass energies up to $\sqrt{s_{\text{NN}}} = 5\,\text{TeV}$.

2 The quark-gluon plasma

QCD is an *asymptotically free* theory, which means that the coupling decreases as the energy scale increases and the corresponding length scale decreases[2]. It can therefore be expected that when nuclear matter is compressed and/or heated eventually the strong interaction will become so weak that quarks and gluons start propagating as nearly free particles instead of being confined in colour neutral bound states. This state of matter consisting of deconfined quarks and gluons is called the *quark-gluon plasma* (QGP). Another interesting feature of the QGP is that chiral symmetry is restored in the plasma phase.

In a gas of non-interacting particles in thermal equilibrium the pressure is given by the Stefan-Boltzmann law:

$$p(T) = \frac{\pi^2}{90} \left(N_{\text{B}} + \frac{7}{8} N_{\text{F}} \right) T^4 \,, \tag{2}$$

where N_{B} and N_{F} are the number of bosons and fermions, respectively. A simple counting of degrees of freedoms leads to $N_{\text{B}}^{(\text{QGP})} = 2[\text{spin}] \cdot (N_c^2 - 1)[\text{colour}] = 16$ and $N_{\text{F}}^{(\text{QGP})} = 2[\text{anti-/particle}] \cdot 2[\text{spin}] \cdot N_c[\text{colour}] \cdot N_f[\text{flavour}] = 36$ for the QGP, while in a hadron gas with temperatures between the pion and the ρ mass one has $N_{\text{B}}^{(\text{hg})} = 3$ and $N_{\text{F}}^{(\text{hg})} = 0$. One thus expects a strong increase of the pressure at the transition from the hadronic phase to the QGP.

Thermodynamic properties of strongly interacting matter can be obtained from Lattice QCD, i.e. by solving QCD numerically on a discrete space-time lattice. However, this works reliably only for vanishing baryon chemical potential, i.e. for systems with vanishing net baryon number. For this case Lattice QCD indeed finds a rapid increase of the pressure and a cross-over into the plasma phase at a pseudo-critical temperature $T_{\text{c}} = (154 \pm 9)\,\text{MeV} \approx 1.7 \cdot 10^{12}\,\text{K}$ [3], which corresponds to roughly 10^5 times the temperature in the centre of the sun. Furthermore, the Lattice results show that the so-called trace anomaly $(\epsilon - 3p)/T^4$, which vanishes for an ideal gas, is still non-zero at $T = 400\,\text{MeV}$ and beyond [3]. This suggests that the QGP is strongly coupled even at temperatures well above the

[2] This is caused by the “anti-screening” of the colour charge due to the gluon self-interaction.

pseudo-critical temperature.

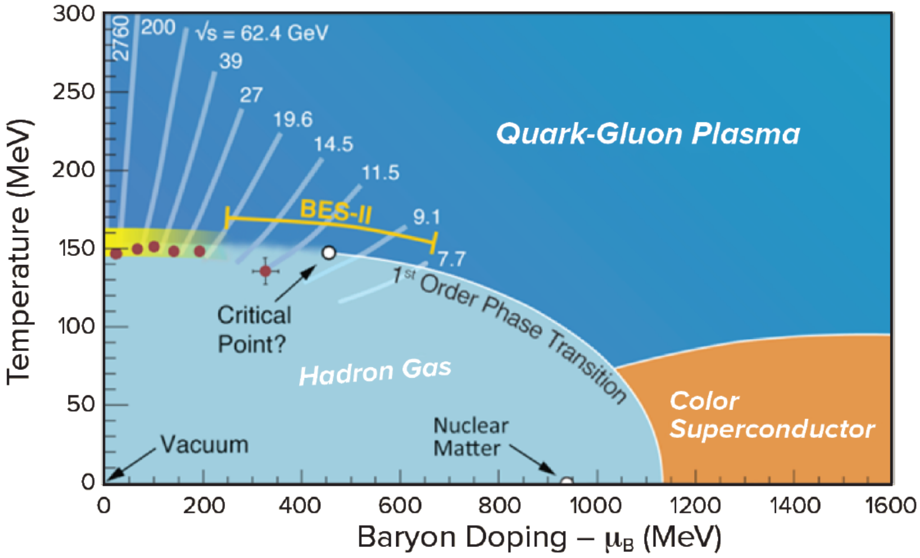


Fig. 1: The phase diagram of QCD with experimental points shown in red, the conjectured first-order phase transition line ending in the critical point and the regions probed by experiments at different beam energies indicated by the light blue lines. The Lattice QCD result for the cross-over temperature at small baryon chemical potential $\mu_{\rm B}$ is shown as the yellow band. Figure from [4].

Figure 1 shows the phase diagram of QCD with baryon chemical potential $\mu_{\rm B}$ on the horizontal and temperature on the vertical axis. Ordinary hadronic matter is found at low temperature T and not too high $\mu_{\rm B}$ and the QGP at high temperature. At high baryon density a colour superconducting phase is expected, but very little is known about this region. The experimentally found points for the phase boundary are shown as the red points. They are extracted from the measured abundances of different hadron species using the *statistical hadronisation model*. The basic assumption of this model is that the system is in thermal equilibrium when it hadronises. The different hadron species are then produced with different probabilities dictated mainly by their mass. It is argued that the composition of the hadronic system does not change afterwards because the density is already too low for number changing scattering processes to occur in the hadronic phase. One can then calculate the expected hadron yields for a grand canonical ensemble and fit the data to extract T and $\mu_{\rm B}$ at the phase boundary. An example for a statistical hadronisation fit is shown in the left plot of figure 2, while the right plot displays the obtained points in the phase diagram. The points at low $\mu_{\rm B}$ are consistent with the Lattice QCD result for the pseudo-critical temperature. These points come from the highest beam energies. The net baryon density in the central rapidity region decreases with beam energy, because the nuclei become more and more transparent. At the LHC the chemical potential extracted from the statistical model is consistent with zero. Another point is remarkable, namely that the statistical hadronisation fits find that strangeness is not suppressed at top RHIC and LHC energies. This indicates that the temperature in the QGP is high enough that strange quarks can be produced thermally. Strangeness enhancement, or rather the disappearance of strangeness suppression, has long been regarded as a sign of QGP formation.

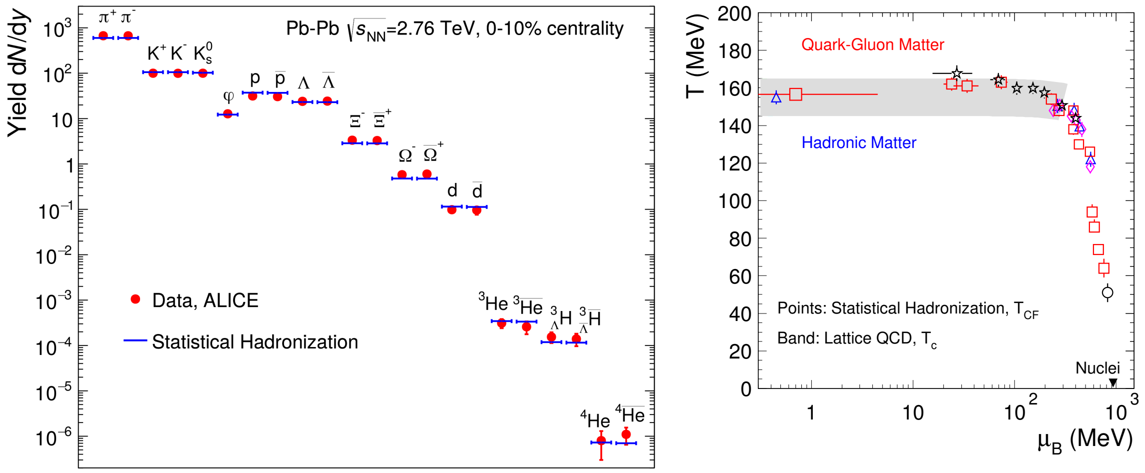


Fig. 2: Left: Hadron yields measured by the ALICE experiment in Pb+Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 2.76\,\mathrm{TeV}$ and statistical hadronisation fit. **Right:** Statistical hadronisation points for different beam energies in the phase diagram. Figures from [9].

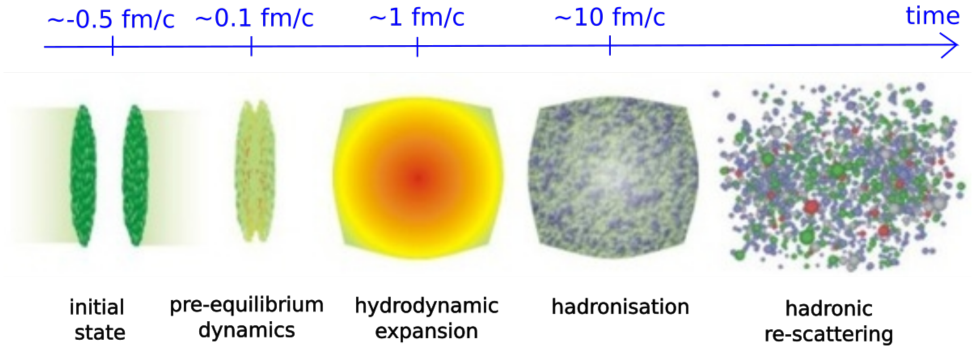


Fig. 3: Different stages of a heavy ion collision at top RHIC and LHC energies, see text for details. Figure courtesy to S. Bass.

Figure 3 summarises the current understanding of heavy ion collisions (see [5–8] for reviews). The evolution of the system produced in a heavy ion collision at high energies encompasses different stages. Starting from the initial state before the actual collision, the system shortly after the collision, at around 0.1 fm/c, enters a phase of pre-equilibrium dynamics that rapidly equilibrates it to a degree that at around 1 fm/c viscous hydrodynamics becomes applicable. It follows an extended phase of hydrodynamic expansion during which the system cools until it reaches the pseudo-critical temperature and hadronises. Re-scattering continues for a while in the hadronic phase until the density becomes too low and the system reaches the kinetic freeze-out. After that the particles free-stream to the detector and resonances decay.

In the following sections I will briefly go through the different stages of a heavy ion collision. After that, I will discuss hard probes, small collision systems and ultra-peripheral collisions.

3 The initial state

When two nuclei collide the amount of overlap can be quantified in different ways. The *impact parameter* b is the transverse distance between the centres of the nuclei. A head-on collision thus has zero impact parameter. The impact parameter is not an observable quantity and inferring it from data inevitably relies on models. The experiments therefore use *centrality*, which is defined as fraction of the geometric cross section. If the nuclei were hard spheres ("billard balls") the geometric cross section would be given by $\sigma_{\text{geo}} = \pi(R_A + R_B)^2$, where R_A and R_B are the radii of the colliding nuclei. The cross section for a collision with impact parameter up to b is then πb^2 and the centrality is the ratio, i.e. $b^2/(R_A + R_B)^2$. The most central collisions ($b = 0$) thus have, somewhat counter-intuitively, zero centrality.

Experimentally, centrality is determined from the event multiplicity and/or the number of forward-going neutrons measured in the zero degree calorimeters (spectator neutrons) [10]. The measured centrality is interpreted in terms of a *Glauber model* [11] to extract the number of nucleons participating the in the collision, N_{part}, and the number of binary nucleon–nucleon collisions, N_{coll}.

Simple versions of the (optical) Glauber model [12] assume that nucleons travel on straight lines, that the nuclei are large enough that edge effects can be neglected, and that there are no correlations among the nucleons inside a nucleus. The nuclei are represented by a smooth density, usually the Woods-Saxon potential

$$n_A(r) = \frac{n_0}{1 + \exp\left(\frac{r-R}{d}\right)} \,. \tag{3}$$

Integrating over the beam direction one obtains the nuclear thickness function

$$T_A(s) = \int\limits_{-\infty}^{\infty} \mathrm{d}z \, n_A(\sqrt{s^2 + z^2}) \,. \tag{4}$$

The overlap between two nuclei colliding with impact parameter b is then found by integrating the product of the two thickness functions

$$T_{AB}(b) = \int \mathrm{d}^2 s T_A(\mathbf{s}) T_B(\mathbf{s} - \mathbf{b}) \,. \tag{5}$$

The number of binary nucleon–nucleon collisions is then obtained by multiplying the thickness function with the inelastic nucleon–nucleon cross section: $N_{\text{coll}} = T_{AB}(b)\,\sigma_{\text{inel}}^{\text{NN}}$. The geometric cross section is then the cross section for having $N_{\text{coll}} \geq 1$. It is given by

$$\sigma_{\text{geo}} = \int \mathrm{d}^2 b \left[1 - e^{-T_{AB}(b)\,\sigma_{\text{inel}}^{\text{NN}}}\right] \,. \tag{6}$$

The number of participating nucleons can be calculated from the probability for a nucleon to pass through nucleus A at impact parameter b without interaction $\mathcal{P}_0^{(A)}(b) = [1 - \sigma_{\text{inel}}^{\text{NN}} T_A(b)/A]^A$ and is given by

$$N_{\text{part}}(b) = \int \mathrm{d}^2 s \, T_A(\mathbf{s}) \left\{1 - \mathcal{P}_0^{(B)}(\mathbf{s} - \mathbf{b})\right\} + \int \mathrm{d}^2 s \, T_B(\mathbf{s}) \left\{1 - \mathcal{P}_0^{(A)}(\mathbf{s} + \mathbf{b})\right\} \,. \tag{7}$$

As a general rule soft (i.e. low $p_\perp$) particle production scales with the number of participants N_{part} while hard (high $p_\perp$) processes are proportional to the number of binary collisions N_{coll}.

Monte Carlo (MC) versions of the Glauber model are a simple way of dealing with the event-by-event fluctuations in the nuclear density. They distribute nucleons inside the nuclei according to the nuclear potential. When the transverse distance between two nucleons from different nuclei is smaller than $\sqrt{\sigma_{\text{inel}}^{\text{NN}}/\pi}$ they are counted has having an interaction. In this way the determination of N_{part} and N_{coll} is straightforward.

Soft particle production probes the gluon density in the nucleons at small x. The scale evolution of the parton distribution functions leads to a rapid rise of the gluon density at small x when Q^2 increases. The DGLAP equations are linear and generate an ever increasing gluon density, but on physical grounds one expects that eventually the density gets so high that gluon recombination becomes important and competes with the splitting processes producing soft gluons. The generic expectation is that this will slow down the evolution and eventually lead to *gluon saturation*. At this point the gluon density cannot increase further. The typical transverse momentum of saturated gluons defines the saturation scale Q_{s}, which can also be regarded as the (transverse) size of saturated gluons. At top RHIC and LHC energies the saturation scale is found to be of the order of a few GeV. This has lead to the development of the *Colour Glass Condensate* (CGC) [13, 14] framework. This picture is valid at high energies when the valence quarks can be approximated as "frozen" by time dilation. They act as sources for saturated gluons with typical momenta $\mathcal{O}(Q_{\text{s}})$. The gluons have occupation number $1/\alpha_{\text{s}}$, i.e. they form an over-occupied state. When the saturation scale is high enough $\alpha_{\text{s}}(Q_{\text{s}}) \ll 1$, which means that the gluons fields are strong but weakly coupled. They can then be described using classical field theory. The gluon fields obey an evolution equation, the so-called JIMWLK equation. The interaction between two nuclei then leads to the strong colour fields decaying to partons.

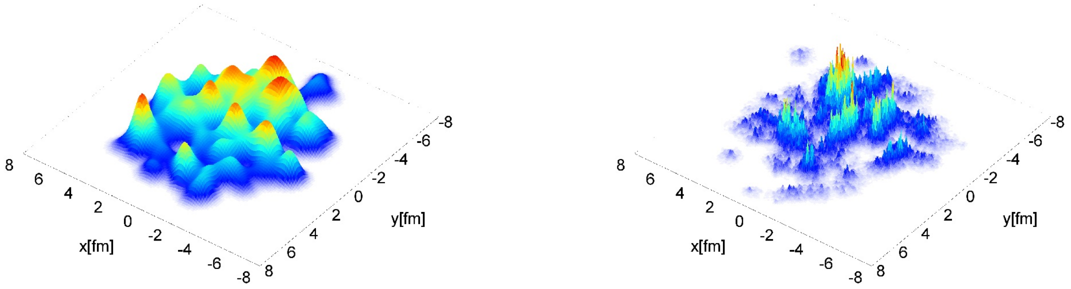


Fig. 4: Energy density in the transverse plane for a Glauber model (left) and the IP-Glasma model encoding gluon saturation in the CGC framework (right), figure from [15].

A characteristic of the CGC is that the typical length scale for fluctuations of the energy density in the initial state is $1/Q_{\text{s}}$. A naive but phenomenological successful alternative approach is to model the energy density by assuming that each participant found in a MC Glauber simulation adds a Gaussian profile of energy density in the transverse plane. The parameters are then found by fitting to data. In case of the calculation shown in figure 4 the width of the Gaussian is $0.4\,\text{fm}$, which is much larger than the $1/Q_{\text{s}}$ of models based on gluon saturation (cf. RHS of figure 4).

4 Pre-equilibrium dynamics

At very early times after the collision of two nuclei the produced system is very far from thermal equilibrium. Despite the rapid longitudinal expansion the system is well described by viscous hydrodynamics at (proper) times of roughly 1 fm/c. The realisation that at that time the system is still very anisotropic has lead to the creation of the term *hydrodynamisation*. Hydrodynamisation denotes the evolution to a point where the system is well described by viscous hydrodynamics (but is not in local thermal equilibrium).

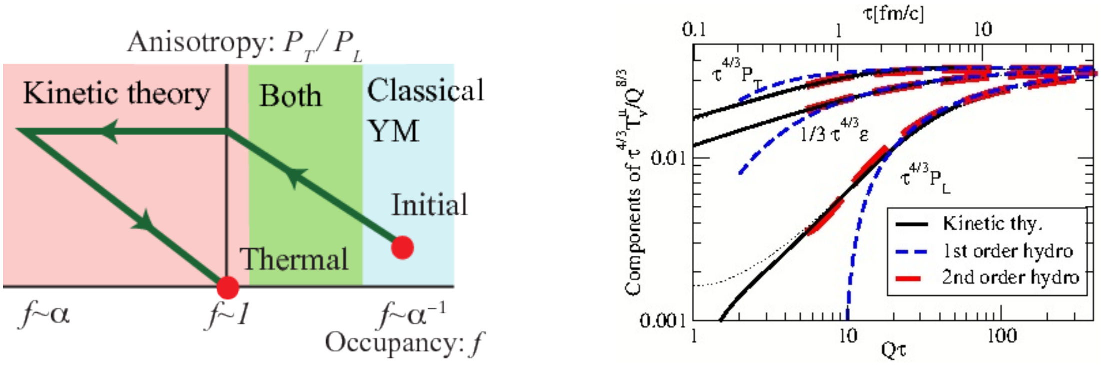


Fig. 5: Left: Cartoon of the evolution from a CGC like initial condition to thermal equilibrium in the presence of strong longitudinal expansion in a weak coupling scenario, figure from [16]. **Right:** Time evolution of rescaled components of the energy-momentum tensor in the effective kinetic theory of QCD at high temperature and hydrodynamics, figure from [17].

The rapid hydrodynamisation can be understood in terms of weak and strong coupling dynamics. In the former scenario the evolution of the system is sketched in figure 5 (left): Starting from a CGC-like initial condition with occupancy $\mathcal{O}(1/\alpha_s)$ the system initially gets more anisotropic and dilute because the interactions are not strong enough to counteract the rapid expansion. Once the systems becomes, under occupied the anisotropy does not increase further. During the last phase the system approaches thermal equilibrium through radiative break-up of the now under occupied modes.

This evolution can be calculated quantitatively using an effective *kinetic theory* of QCD at high temperatures [18]. It formulates a set of Boltzmann equations

$$-(\partial_t + \mathbf{v}\cdot\nabla_x) f(\mathbf{x},\mathbf{p},t) = \mathcal{C}_{``1\leftrightarrow 2"}[f] + \mathcal{C}_{2\leftrightarrow 2}[f] \tag{8}$$

for the phase space densities $f(\mathbf{x},\mathbf{p},t)$ of (anti-)quarks and gluons. The dynamics is encoded in the collision kernels on the right hand side, where $\mathcal{C}_{2\leftrightarrow 2}[f]$ describes elastic scattering and $\mathcal{C}_{``1\leftrightarrow 2"}[f]$ nearly collinear splitting and merging processes in the presence of multiple coherent scattering. Solving the Boltzmann equations numerically shows that the energy-momentum tensor coincides with the expectation from second order viscous hydrodynamics starting from times $\lesssim 1$ fm/c (figure 5 right).

Strong coupling calculations rely in the *AdS/CFT correspondence* relating strongly coupled conformal field theories to weakly coupled type IIB string theories in a five-dimensional AdS space. Most of the results have been obtained for $\mathcal{N} = 4$ Super-Yang-Mills theory, which shares many similarities with QCD at high temperature. Heavy ion collisions are in this scenario usually modeled as collision of two shock waves [19]. Thermalisation of the produced system is related to the creation of a black hole in the fifth dimension of the string theory. Also in this case the system quickly reaches hydrodynamic behaviour on timescales $\mathcal{O}(1/T)$.

5 Hydrodynamics

One of the most compelling pieces of evidence for hydrodynamic behaviour is that the anisotropic flow coefficients are well described by hydrodynamics. What is meant by this is the following: the system created in mid-central collisions has an elliptical shape in the transverse plane. In a hydrodynamic scenario the expansion of the system is driven by pressure gradients. In the case of an elliptically shaped system the pressure gradient is larger along the short axis than along the long axis. As a result, matter is pushed out preferentially in the direction of the short axis, which leads to an isotropic momentum distribution. This is quantified by a Fourier decomposition of the momentum distribution in azimuthal angle ϕ

$$\frac{\mathrm{d}\,N}{\mathrm{d}\,\phi} = \frac{N}{2\pi}\left[1 + 2\sum_n v_n \cos(n(\phi - \Psi_n))\right] , \tag{9}$$

where the event plane angle Ψ gives the orientation of the collision system in azimuth. The coefficients v_n are the so-called *flow coefficients*. The elliptical shape of the overlap region induces a large elliptical momentum anisotropy, i.e. a large v_2. Figure 6 shows a measurement of v_2 compared to a hydrodynamic calculation. The data are very well described by the calculation up to transverse momenta of roughly 2 GeV, beyond which other particle production mechanisms become dominant. The different hadron species show a mass ordering that is characteristic for *collective flow*, where all particles flow with a common velocity.

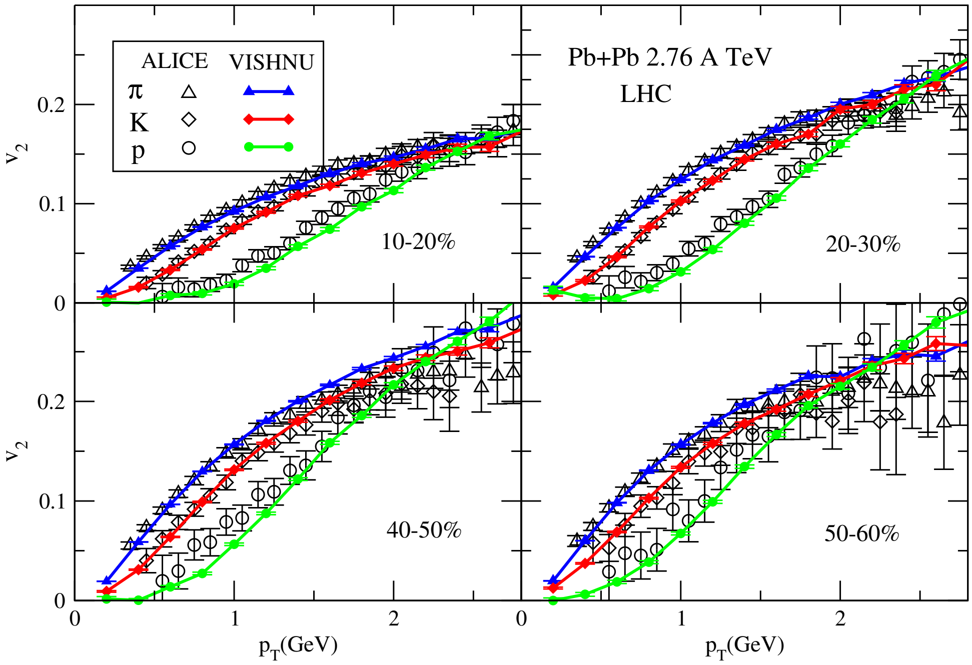


Fig. 6: Measurement of the elliptic flow coefficient v_2 for different hadron species in different collision centralities compared to a hydrodynamic calculation, figure from [20].

Hydrodynamics [21,22] can be viewed as a low energy effective theory describing long distance, late time behaviour of averaged macroscopic features of a system. It is applicable to a very generic set of theories and assumes that the system is in local thermal equilibrium, i.e. a temperature can be defined locally at each point but may vary from point to point. Microscopic dynamics of the theory enter through the equation of state and a set of coefficients. The equation of state relates the energy density ϵ and the pressure p. Hydrodynamics is valid for distances that are long compared to the mean free path, times that are long compared to the inverse scattering rate and systems with sufficiently smooth variations.

A fluid in thermal equilibrium is described by the energy-momentum tensor

$$T^{\mu\nu} = \epsilon u^\mu u^\nu + p(g^{\mu\nu} + u^\mu u^\nu)\,, \tag{10}$$

where $g^{\mu\nu}$ is the metric and u^μ the fluid velocity. We here allow for small deviations from global thermal equilibrium such that energy density and fluid velocity vary with position and time, $\epsilon = \epsilon(x)$ and $u^\mu = u^\mu(x)$. Energy-momentum conservation $\partial_\mu T^{\mu\nu} = 0$ leads to

$$u^\mu \partial_\mu \epsilon + (\epsilon + p)\partial_\mu u^\mu = 0 \tag{11}$$

$$(\epsilon + p)u^\mu \partial_\mu u^\nu + (g^{\nu\mu} + u^\nu u^\mu)\partial_\mu p = 0\,. \tag{12}$$

Together with the equation of state these form a closed system that can be solved to obtain the energy-momentum tensor.

In order to allow for perturbations with larger gradients a more general form of the energy-momentum tensor is needed

$$T^{\mu\nu} = \epsilon u^\mu u^\nu + p\Delta^{\mu\nu} + \Pi^{\mu\nu}\,, \tag{13}$$

where $\Delta^{\mu\nu} = g^{\mu\nu} + u^\mu u^\nu$ and $\Pi^{\mu\nu}$ is the viscous stress tensor. It can be decomposed into a traceless part and the remainder

$$\Pi^{\mu\nu} = \pi^{\mu\nu} + \pi_{\text{bulk}}\Delta^{\mu\nu}\,, \tag{14}$$

where the shear stress tensor $\pi^{\mu\nu}$ and the bulk viscous pressure π_{bulk} parametrise deviations from ideal fluid dynamics. Viscous fluid dynamics can be organised as a gradient expansion

$$\pi_{\text{bulk}} = -\zeta \partial_\mu u^\mu + \dots \tag{15}$$

$$\pi^{\mu\nu} = -2\eta \left(\frac{1}{2}\Delta^{\mu\alpha}\Delta^{\nu\beta} + \frac{1}{2}\Delta^{\mu\beta}\Delta^{\nu\alpha} + \frac{1}{3}\Delta^{\mu\nu}\Delta^{\alpha\beta} \right) \partial_\alpha u_\beta + \dots \tag{16}$$

where at first order the bulk viscosity $\zeta = \zeta(\epsilon)$ and the shear viscosity $\eta = \eta(\epsilon)$ appear. The second order comes with many more parameters (relaxation times, more transport coefficients, ...). The resulting increasingly complicated evolution equations have to be solved numerically except for the simplest cases.

The shear viscosity is related to momentum transport in the fluid. When the shear viscosity is large compared to the entropy density (large η/s) momentum perturbations are transported over large distances by *quasi-particles*, while at small η/s there are no well-defined quasi-particles. The shear viscosity thus also determines to what extend perturbations in the initial conditions are washed out. This can be seen in figure 7, which shows how an initial energy density with fluctuations looks after evolving

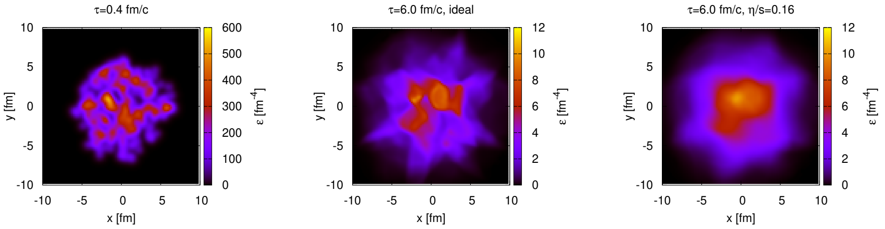


Fig. 7: Initial energy density distribution at proper time $\tau = 0.4\,\mathrm{fm/c}$ in the transverse plane (left) and after evolving to time $\tau = 0.6\,\mathrm{fm/c}$ with ideal hydrodynamics (middle) and with viscous hydrodynamics with a small shear viscosity $\eta/s = 0.16$, figure from [23].

it for a short time with and without a small shear viscosity. With shear viscosity the fluctuations have been smeared out to a larger extent.

The efficiency with which elliptic flow is generated from a given spatial anisotropy also depends on the shear viscosity — the smaller the shear viscosity the larger v_2 [24]. The values of η/s that have been inferred from measurements of the flow coefficients are very close to $1/(4\pi)$, which is a conjectured lower bound obtained from AdS/CFT calculations [25]. The QGP is thus the least dissipative material currently known.

Given the shape of the overlap region of two colliding nuclei the elliptic flow coefficient v_2 is dominant, but higher coefficients are also present. While a smooth spatial distribution with a mirror symmetry does not generate odd harmonics ($n = 3, 5, \dots$) event-by-event fluctuations in the initial state give rise to these.

The measurements and the hydrodynamic modeling have become so precise that they are sensitive to small deformations of the initial state, e.g. due to deformations of the colliding nuclei. This is for instance the case in data from Ru+Ru and Zr+Zr collisions at RHIC [27].

6 Hadronisation

Hydrodynamics describes the evolution of a fluid, but in the detector particles are measured. The usual way of translating the fluid's energy-momentum tensor into a particle spectrum is the *Cooper-Frye prescription* [28]. The starting point is the observation that, neglecting viscous corrections, the occupation number in each fluid element is given by a Fermi-Dirac or Bose-Einstein distribution

$$\frac{\mathrm{d}N_i}{\mathrm{d}^3x\,\mathrm{d}^3p} = f_i(p^\mu; T(x), u^\mu(x)) \propto \left[e^{\frac{u_\mu(x)p^\mu}{T(x)}} \pm 1\right]^{-1} . \tag{17}$$

To find the particle spectrum one considers the number of particles passing through a surface Σ, which can be written in terms of the particle number current density $j_i^\mu(x)$ for particle species i

$$N_i = \int_\Sigma \mathrm{d}^3\sigma_\mu\, j_i^\mu(x) = \int_\Sigma \mathrm{d}^3\sigma_\mu \left[\int \frac{\mathrm{d}^3p}{(2\pi)^3}\,\frac{p^\mu}{E} f_i(x,p)\right] , \tag{18}$$

where $\mathrm{d}\sigma_\mu$ is the area element of the surface Σ. The particle spectrum is thus found to be given by

$$E\frac{\mathrm{d}N_i}{\mathrm{d}^3p} = \frac{1}{(2\pi)^3}\int\limits_\Sigma p_\mu \mathrm{d}\sigma^\mu \, f_i(x,p)\,. \tag{19}$$

The momentum spectrum of particles obtained from two different surfaces is the same only if the particles are free-streaming between the two surfaces. The obvious choice for Σ is thus the freeze-out surface, i.e. the surface of last scattering. In practice it is often assumed that the temperature is constant on the freeze-out surface.

When the partonic distribution function is known the distribution of hadrons can be obtained from the *quark coalescence* picture [29,30]. The idea of quark coalescence is simply put that quarks and antiquarks combine to form hadrons. The number of mesons, for instance, is then again given by an integral over a space-like surface Σ

$$N_\mathrm{M} = g_\mathrm{M}\int_\Sigma (p_{1\mu}\mathrm{d}\sigma_1^\mu)\,(p_{2\mu}\mathrm{d}\sigma_2^\mu)\,\frac{\mathrm{d}^3p_1}{(2\pi)^3E_1}\,\frac{\mathrm{d}^3p_2}{(2\pi)^3E_2} f_q(x_1,p_1)f_{\bar{q}}(x_2,p_2)f_\mathrm{M}(x_1,x_2,p_1,p_2)\,, \tag{20}$$

where g_M is a statistical factor and f_M is the probability for a quark and anti-quark to form a meson, e.g.

$$f_\mathrm{M}(x_1,x_2,p_1,p_2) \propto \exp\left(\frac{(x_1-x_2)^2}{2\Delta_x}\right)\cdot \exp\left(\frac{(p_1-p_2)^2}{2\Delta_p}\right)\,. \tag{21}$$

For baryons corresponding expressions can be written down.

The coalescence picture predicts that v_2 of hadrons scales with the number of constituent quarks, a feature that is clearly visible in RHIC data (and to a lesser extent in LHC data).

7 Hadronic re-scattering

After the chemical freeze-out where quarks and gluons are converted into hadrons re-scattering continues for a while in the hadronic phase. There are two ways of dealing with this. The first is to run hydrodynamics down to the kinetic freeze-out, which requires additional input to adequately describe the hadronic phase. The second option is to explicitly simulate re-scattering in the hadronic phase with transport codes based on the Boltzmann equation. This requires knowledge of a large number of hadronic resonances and cross sections. Examples for hadronic transport codes are UrQMD [31] and SMASH [32].

8 Hard probes

Particles with very high transverse momentum are produced in hard partonic scattering processes characterised by a large momentum transfer Q. Due to the uncertainty principle these processes occur on a timescale $\mathcal{O}(1/Q)$ and are therefore the first processes to happen in a collision of two nuclei. The produced high-$p_\perp$ particles then travel through the QGP on their way to the detector. The time and length scale of hard scattering processes are too short to feel the nuclear environment. This is basically the same argument as underlying factorisation theorems that allow to write the cross section for a hard process as a convolution of a partonic cross section $\hat{\sigma}_{ij}$ for producing the hard particle in a scattering of partons i

and j with the parton distribution functions $f_i(x, Q^2)$

$$\sigma(P_1, P_2) = \sum_{i,j} \int_0^1 \mathrm{d}x_1\, \mathrm{d}x_2\, f_i(x_1, Q^2) f_j(x_2, Q^2) \hat{\sigma}_{ij}(x_1 P_1, x_2 P_2, \alpha_\mathrm{s}, Q^2)\,. \tag{22}$$

Here the partonic cross section $\hat{\sigma}_{ij}$ encodes the short distance physics and is insensitive to the nature of the colliding hadrons. It has an expansion in powers of the strong coupling α_s and can be calculated order by order in perturbation theory. The parton distribution function (PDF) $f_i(x, Q^2)$ is the density of partons of type i carrying a fraction x of the hadron's longitudinal momentum inside the beam hadron when probed at a scale Q^2. The PDFs of protons and neutron bound inside nuclei are somewhat different from the free proton PDF and there are *nuclear PDF* fits available.

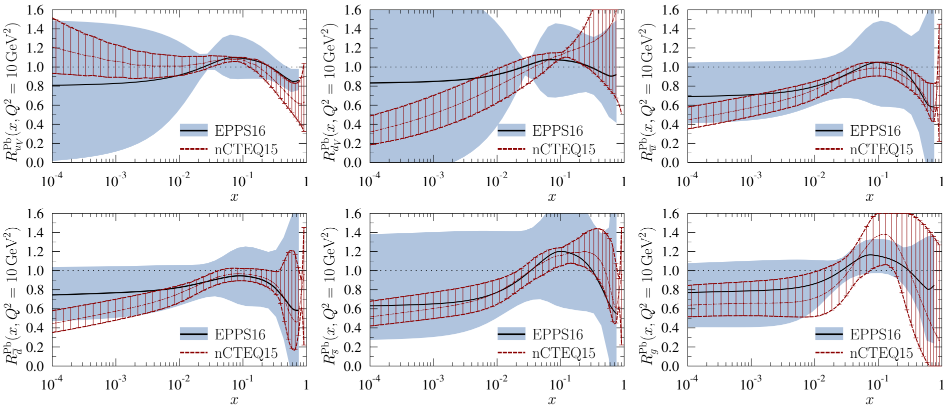


Fig. 8: Nuclear modifications of the PDFs for Pb nuclei at a scale of $Q^2 = 10\,\mathrm{GeV}^2$ obtained from the EPPS16 and nCTEQ15 nuclear PDF sets, figure from [33].

The approach taken by the EPPS collaboration [33] is to define the bound proton PDF as

$$f_i^{\mathrm{p/A}}(x, Q^2) = R_i^A(x, Q^2) f_i^{\mathrm{p}}(x, Q^2)\,, \tag{23}$$

where f_i^{p} is a free proton PDF and R_i^A encodes the nuclear modification. The free proton PDF used for the EPPS16 nuclear PDF set is CT14NLO. R_i^A is parametrised and fitted to data in a procedure that is similar to the way in which free proton PDFs are constructed. The reason for not directly fitting the entire nuclear PDF (instead of the nuclear modification of the free proton PDF) is that the amount of data available for nuclear PDF fits is much smaller than for free proton PDFs. The neutron PDF is obtained from isospin symmetry. Figure 8 shows a comparison of the nuclear modification of the PDF from the EPPS16 [33] and nCTEQ15 [34] nuclear PDF sets. Other nuclear PDF sets, e.g. DSSZ [35], are also available but not shown. Generally, the uncertainties are sizable due to sparse data from nuclear collisions. The nuclear modifications are typically of moderate size and become smaller at higher scales. From a theoretical standpoint the production of high-$p_\perp$ particles in hard scattering processes is thus under control.

The production of the heavy charm and beauty quarks requires a high momentum scale due to

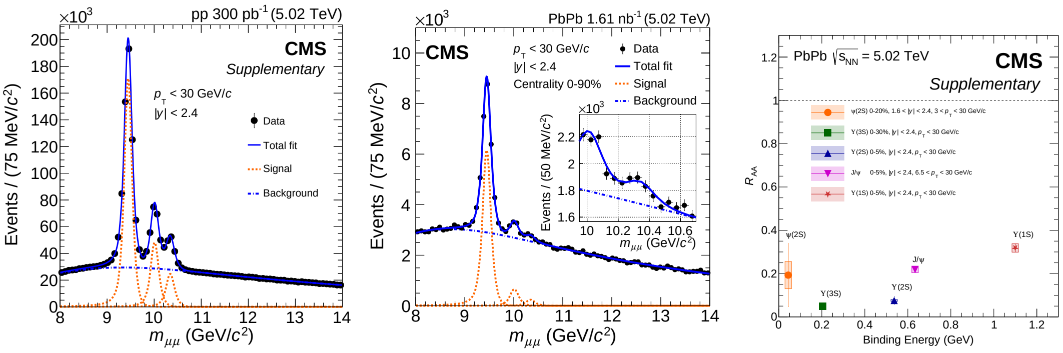


Fig. 9: Left and middle: Invariant mass distribution of muon pairs showing the $\Upsilon(1S)$, $\Upsilon(2S)$ and $\Upsilon(3S)$ peaks in p+p (left) and Pb+Pb (middle) collisions as measured by CMS. **Right:** Quarkonium yields per binary collision in Pb+Pb collision divided by the respective p+p yields showing the suppression of quarkonia states in nuclear collisions, figures from [37].

Table 1: Masses, binding energies and radii of the lightest charmonium and bottomonium states.

state	J/Ψ	Ψ′	Υ	Υ′	Υ″
mass [GeV]	3.10	3.68	9.46	10.02	10.36
ΔE [GeV]	0.64	0.05	1.10	0.54	0.20
r [fm]	0.25	0.45	0.14	0.28	0.39

the large masses of the produced particles. An effect that has long been regarded as a signature for QGP formation is the suppression of quarkonium states due to *colour screening* [36]. The idea is that when the quarkonium states are placed in a deconfined medium consisting of quasi-free colour charges the charges of the heavy quarks are screened. When the screening length is smaller than the size of the quarkonium state the quarkonium is expected to dissolve. Different quarkonium states have different radii and should thus disappear at different QGP temperatures. Observing this so-called sequential suppression should thus in principle allow to measure the QGP temperature. Table 1 lists the binding energies and radii of the most easily accessible quarkonium states. As seen in Figure 9 there is indeed an indication of a binding energy dependent suppression of quarkonium states in heavy ion collisions.

There is, however, a competing effect and that is the *regeneration* of quarkonia by statistical hadronisation at the phase boundary [38]. At the LHC the charm cross section is so large that typically more than one $c\bar{c}$ pair is produced per Pb+Pb collision. When the number of charm quarks in the system is large enough there is a sizable probability that a c and a $\bar{c}$ quark will end up close in phase space and form a charmonium state during hadronisation. This effect can even lead to an enhancement of charmonium states.

Hard processes are expected to scale with the number of binary collisions, therefore a common way of quantifying modifications due to a nuclear environment is to take the ratio to the corresponding quantity in p+p collisions scaled by the number of binary collisions

$$R_{AA}(\{\cdot\}) = \frac{\left.\frac{\mathrm{d}\sigma_X}{\mathrm{d}\{\cdot\}}\right|_{AA}}{\langle N_{\mathrm{coll}}\rangle \left.\frac{\mathrm{d}\sigma_X}{\mathrm{d}\{\cdot\}}\right|_{pp}} . \tag{24}$$

Quantities of this type are called *nuclear modification factors*. In the case of electroweak processes the different isospin composition in p+p and heavy ion collisions also has to be taken into account. Deviations from unity in nuclear modification factors are signs that a nucleus–nucleus collision is more than the incoherent superposition of nucleon–nucleon collisions and that there are additional effects from the nuclear environment.

The electroweak bosons W, Z and γ are produced in hard scattering processes but don't participate in the strong interaction and thus escape from the QGP without interaction. Measurements of nuclear modification factors for these bosons are indeed consistent with the expectation from nuclear PDF and isospin effects [39, 40]. This is an important confirmation of scaling with N_{coll} and a cross-check that nuclear PDF effects are under control.

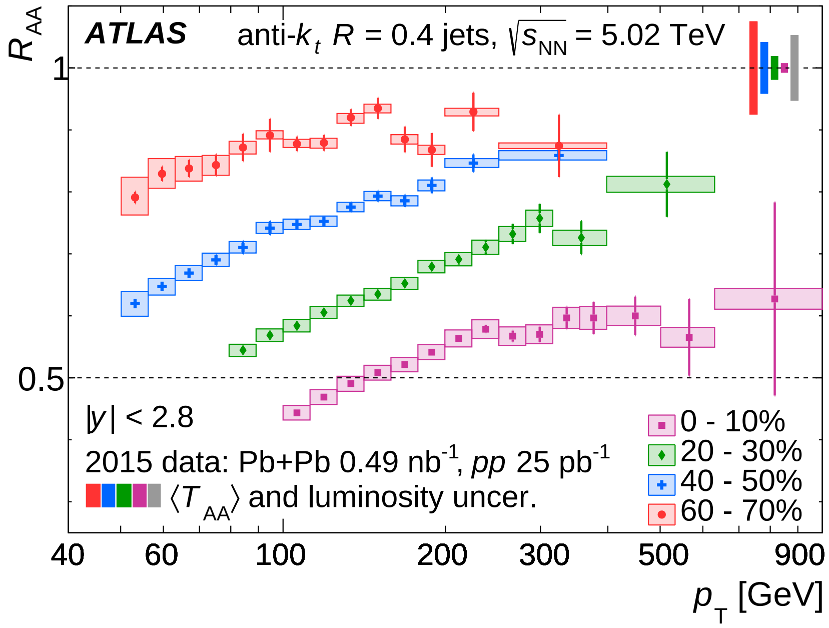


Fig. 10: Nuclear modification factor for jets as a function of the jet $p_\perp$ in Pb+Pb collisions of different centrality as measured by ATLAS, figure from [41].

The situation is very different for jets, which are the result of the fragmentation of highly energetic quarks and gluons. Figure 10 shows that jets are suppressed by almost a factor of two out to very high $p_\perp$ in central Pb+Pb collisions at the LHC. This phenomenon (together with modifications of the internal structure of jets) is referred to as *jet quenching*. The energetic quarks and gluons interact in the QGP and thereby loose energy. Jet quenching can thus be seen as the partial thermalisation of a far-from-equilibrium system and can inform us about equilibration in QCD.

Another observation is that the transverse momentum imbalance in di-jets increases, because one of the jets looses more energy than the other [42]. Interestingly, the azimuthal angle between the two jets remains unchanged. Looking at the energy distribution inside jets it is found that the jet profile shows a suppression at intermediate distances from the jet axis and an enhancement at the periphery of the jet, i.e. at large distances from the jet axis [43]. The jet fragmentation functions show a consistent enhancement of soft fragments and a suppression at intermediate momenta [44]. One can thus conclude that hard structures inside jets survive largely unmodified while there is a clear enhancement of soft activity at large angles relative to the jet axis. This indicates that hard partonic systems loose energy predominantly through radiation of soft gluons while rare hard or semi-hard emissions don't play a major role [45].

There are two scenarios for how hard partons interact in the QGP, namely the strong coupling and the weak coupling scenario. In the former the hard parton does not resolve quasi-particles in the QGP and one has to employ AdS/CFT techniques to calculate the energy loss of hard partons at strong coupling, which is caused by a kind of drag force. The advantage of this approach is that an exact solution can be obtained with no uncertainties in the parton-medium interaction. A fundamental problem is that jets are a weak coupling phenomenon and don't have a natural counterpart at strong coupling. A $q\bar{q}$ pair in a QGP is dual to a classical string falling into a black hole in the fifth dimension of AdS space. This construction has been used as a proxy for a "holographic jet" [46, 47].

In the weak coupling scenario hard partons scatter off quasi-particles in the QGP. It should be noted that the relevant scale for distinguishing between the two scenarios is the momentum transfer between the hard parton and the QGP (not the momentum of the hard parton). Perturbative techniques can be used to calculate the energy loss of a hard parton in QCD at week coupling. Two types of processes can occur, namely elastic scattering and bremsstrahlung. Thermalisation through elastic scattering is slow and the energy loss is dominated by QCD bremsstrahlung. In bremsstrahlung due to multiple scattering a quantum mechanical interference occurs because scatterings within the formation time of a gluon emission act coherently. This gives rise to the QCD analogue of the Landau-Pomerantchuk-Migdal (LPM) effect.

The concept of *formation time*, i.e. the time it takes to radiate a gluon, plays a central role in the discussion of energy loss of hard partons. The emission of a gluon at finite emission angle is kinematically only possible when the emitting parton is off-shell with a virtual mass $m_\text{virt} = p^2$. The time it takes the virtual parton with energy E to "decay" by emitting a gluon with energy ω and relative transverse momentum $k_\perp$ is the gluon formation time and can be estimated from the uncertainty principle to be $\mathcal{O}(1/m_\text{virt})$ in the virtual parton's rest frame. In a general frame the formation time acquires an additional boost factor E/m_virt and becomes

$$t_\text{form} = \frac{1}{m_\text{virt}} \frac{E}{m_\text{virt}} = \frac{E}{2p_\mu k^\mu} \simeq \frac{E}{\omega E \theta^2} \simeq \frac{\omega}{k_\perp^2} \,, \tag{25}$$

where p_μ and k_μ are the parton and gluon momentum, respectively, and θ is the emission angle.

To illustrate the LPM effect [48], let's consider a highly energetic quark with energy E traveling through a coloured background medium and emitting an almost collinear gluon with energy $\omega \ll E$. Due to multiple scattering in the background the gluon undergoes Brownian motion and acquires a transverse momentum relative to the quark $\langle k_\perp^2 \rangle = \hat{q}L$, where L is the path length traveled and $\hat{q}$ is the transport parameter. Then the gluon formation time becomes

$$t_\mathrm{f} \simeq \frac{\omega}{k_\perp^2} \simeq \frac{\omega}{\hat{q}t_\mathrm{f}} \quad \Rightarrow \quad t_\mathrm{f} = \sqrt{\frac{\omega}{\hat{q}}}\,. \tag{26}$$

The number of scatterings during the formation time is given by $N_\mathrm{coh} = t_\mathrm{f}/\lambda$, where λ is the mean free path. The incoherent gluon spectrum is parametrically $\mathrm{d}^2 I^\mathrm{incoh}/\mathrm{d}\omega\mathrm{d}y \propto \alpha_\mathrm{s}/(\omega\lambda)$, where y is the longitudinal spatial coordinate. To estimate the gluon energy spectrum including coherence effects one has to divide by the number of scatterings acting coherently, i.e. effectively as one

$$\frac{\mathrm{d}^2 I^\mathrm{coh}}{\mathrm{d}\omega\mathrm{d}y} \simeq \frac{1}{N_\mathrm{coh}}\frac{\mathrm{d}^2 I^\mathrm{incoh}}{\mathrm{d}\omega\mathrm{d}y} \propto \frac{\alpha_\mathrm{s}}{\omega\lambda}\lambda\sqrt{\frac{\hat{q}}{\omega}} = \alpha_\mathrm{s}\sqrt{\hat{q}}\,\omega^{-3/2}\,. \tag{27}$$

To obtain the total energy loss this spectrum has to be weighted with gluon energy and integrated over the gluon energy and the path traveled in the medium

$$\Delta E = \int\limits_0^L \mathrm{d}y \int\limits_0^{\omega_\mathrm{c}} \mathrm{d}\omega\, \omega \frac{\mathrm{d}^2 I}{\mathrm{d}\omega\mathrm{d}y} \propto \alpha_\mathrm{s}\hat{q}L^2\,, \tag{28}$$

where $\omega_\mathrm{c} = \hat{q}L^2$ is the highest gluon energy that can be radiated coherently, i.e. the gluon energy for which the formation time equals the path length L in the medium and the entire medium acts coherently. The L^2 dependence of the energy loss is characteristic of the LPM effect.

A high-$p_\perp$ parton produced in a hard scattering process undergoes a scale evolution in QCD during which it emits gluons[3]. Inserting numbers for typical jets at the LHC into equation (25) one finds that the first few emission happen very quickly, usually before the QGP forms. However, the time needed for the total scale evolution through successive gluon emissions is of the order of a few fm/c and thus comparable to the transverse size of the QGP. Interactions in the QGP and gluon emission related to the QCD scale evolution thus happen at the same time.

When a quark emits a gluon the resulting quark and gluon together still form a colour triplet. Whether the quark–gluon system is resolved as a quark carrying a triplet charge and a gluon carrying an octet charge, or is seen as a quark with a triplet charge depends on the resolution scale. For interactions in the QGP this means that a scattering with a transverse momentum transfer larger than the inverse transverse separation of the quark and the gluon will resolve two partons. A soft scattering with transverse momentum transfer smaller than the inverse transverse separation, on the other hand, will only 'see' a quark [49]. This is a type of *colour coherence* and can naturally explain why hard small angle structures inside jets are not affected by jet quenching. The hope that by measuring up to which opening angle structures stay coherent to get a handle on the medium resolution power has so far not been fulfilled.

Many jet quenching measurements and particularly those targeted at the internal structure of quenched jets suffer from the so-called *selection bias*. The jet $p_\perp$ spectrum is very steeply falling and

[3] This is basically the same as the DGLAP evolution of the PDFs and is simulated by parton showers in Monte Carlo event generators or taken care of by analytical resummation.

the energy loss has large fluctuations. As a consequence, in a sample of jets selected based on the final $p_\perp$ those that lost only very little energy will always dominate. The nuclear modification factor is thus mostly sensitive to the no-quenching probability. In jet shape and jet sub-structure observables a related effect becomes visible. As mentioned earlier, the first few splittings of a hard parton happen very early and can thus be expected to be unmodified by the QGP. But they are decisive in determining the general shape of the jet. Already in p+p collisions jets with the same $p_\perp$ are not all the same: some have a soft fragmentation pattern where the energy is shared among many particles already at parton level, while some have a hard fragmentation pattern with only few energetic partons. The soft fragmenting jets are more susceptible to medium modifications and typically loose significantly more energy than the hard fragmenting ones. In the presence of energy loss the soft fragmenting jets are thus more likely to fall below the minimum $p_\perp$ required and thereby disappear from the sample. A sample of jets in heavy ion collisions is thus biased towards a harder fragmentation pattern with a harder and narrower core than in p+p collisions. This is indeed observed in data. This bias is hard to control on a quantitative level and complicates the interpretation of the internal structure of quenched jets.

The QGP modifies hard partons, and hard partons modify the QGP. The energy and momentum lost by hard partons is transferred into the QGP and manifests itself in the form of additional soft particle production at large angles relative to the jet axis [50]. The measurements go out to $\Delta r = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2} = 1$, but most theoretical calculations are available only for smaller $\Delta r < 0.3$. In that regime the enhancement of soft particles is in some calculations clearly caused by the so-called *medium response* [51, 52], in others it is due to other effects [53] or it is a mixture [54]. The situation is thus so far inconclusive.

9 Small collision systems

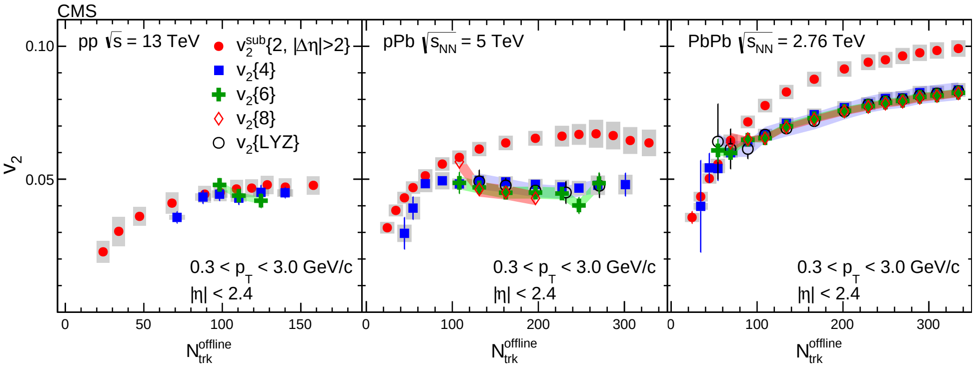


Fig. 11: Elliptic flow coefficient v_2 measured using different methods in p+p, p+Pb and Pb+Pb collisions, figure from [55].

The traditional heavy ion physicists' view was that many correlations observed among soft particles produced in heavy ion collisions are signs of collective flow in the final state and thus indicative of QGP formation. In p+p collisions, on the other hand, the density of produced particles was thought to be too low for sizable final state re-scattering. The expectation was therefore that the particles produced in p+p and other small collision systems like p+Pb free-stream to the detector. It therefore came as a surprise when the LHC experiments discovered that many of the characteristics of heavy ion collisions are also found in small collision systems. An example is v_2, which is shown in figure 11 for p+p, p+Pb and Pb+Pb collisions. Although the magnitude of v_2 increases with system size the overall behaviour is similar in all three systems. Another example is strangeness enhancement, which shows a smooth increase with multiplicity from p+p via p+Pb to Pb+Pb collisions [56]. Interestingly, no jet quenching has so far been observed in small collision systems.

Different explanations for the observed correlations in small collision systems have been put forward. The first is that it is due to response to the initial geometry. The mechanism that creates the azimuthal anisotropy could then be hydrodynamic flow as in nucleus–nucleus collisions [57], or it could be the so-called *escape mechanism* [58, 61] which requires only $\mathcal{O}(1)$ scattering per particle. Alternatively, correlations present in the initial state could get imprinted on the final state without final state interactions. The CGC framework, for instance, predicts such initial state correlations [59]. A third explanation is inspired by p+p physics and proposes string interactions, in this case string shoving, as an explanation for the observed correlations [60]. Presently, no preference for one of these scenarios can be identified.

Despite the observed similarities there are also fundamental differences between small and large collision systems. One is that in small systems multiplicity is generated to a larger extent by jets than in large systems. As a consequence there is no strong correlation between centrality and multiplicity as in heavy ion collisions [62]. Small systems are also affected by auto-correlations of different kinds that can complicate making meaningful measurements. One example for this is that the charged and neutral kaon multiplicities have a different dependence on the charged particle multiplicity. The reason is simply that by requiring a large charged particle multiplicity one biases the events towards having more charged than neutral kaons. Similarly, a rapidity shift of a central di-jet system depending on the amount of transverse energy in the forward region [63] could be explained by energy conservation [64]. These complications are part of the reason why no consistent understanding of small collision systems has emerged so far.

10 Ultra-peripheral collisions

Ultra-relativistic ions are sources of very strong electromagnetic fields. Electric fields can go up to $10^{16} - 10^{18}$ V/m and magnetic fields reach $10^{14} - 10^{16}$ T. These are the strongest electromagnetic fields in the universe, but they are extremely short lived. In ultra-peripheral collisions where the impact parameter is too large for strong interactions to occur a new class of processes with incoming photons becomes accessible. These processes can be roughly divided into photon–photon scattering where two photons, one from each nucleus, interact and photo-nuclear reactions, where a photon from one nucleus interacts with the other nucleus. An example for the first class is the first evidence for light-by-light scattering reported by ATLAS (figure 12). The scattering $\gamma + \gamma \to \gamma + \gamma$ proceeds in the Standard Model via a lepton loop as shown in figure 12. In certain BSM extensions, however, it can proceed via an axion-like particle in the s-channel and the measurements can be used to set limits on the axion-like particles [66].

Both CMS and ATLAS have measured the anomalous magnetic moment a_τ of the τ in $\gamma + \gamma \to \tau + \bar{\tau}$ events [67, 68]. Due to the large τ mass a_τ is 280 times more sensitive to BSM physics than

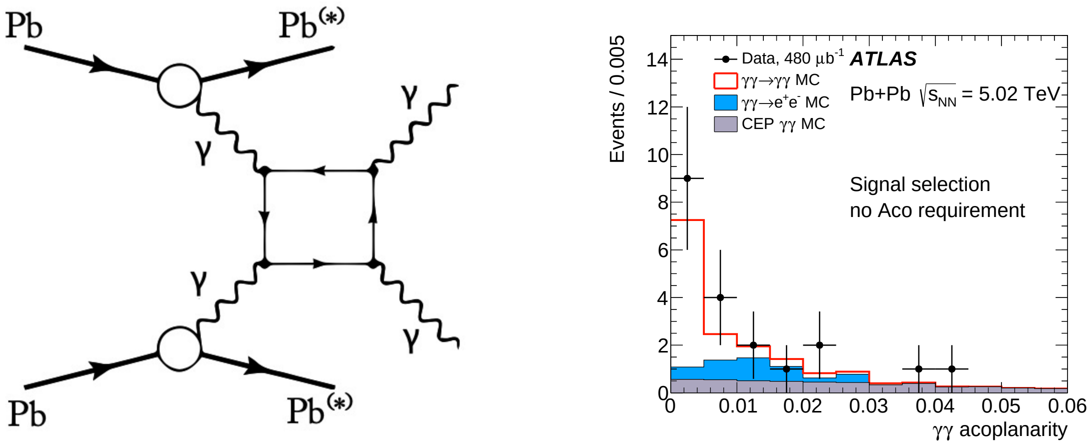


Fig. 12: First evidence for light-by-light scattering in ultra-peripheral Pb+Pb collisions by ATLAS, figure from [65].

the anomalous magnetic moment of the muon. With better statistics expected in the near future the measurements will reach similar precision as the currently best determinations of a_τ.

A similar process, namely the Breit-Wheeler process $\gamma + \gamma \to e^+ + e^-$ can be used to measure the photon spin via the angular asymmetry of the produced leptons [69]. These measurements can therefore be used to place limits on dark photons [70].

An example for an interesting photo-nuclear reaction is diffractive J/Ψ production. By measuring the coherent (where the photon interacts with the whole nucleus) and incoherent (where the photon resolves the nuclear structure) production modes one is sensitive to the nuclear and sub-nucleonic structure, respectively [71, 72].

These are just a few examples for the kind of physics that ultra-peripheral collisions give access to. With larger data sets becoming available this area is certainly going to expand.

11 Conclusions

Heavy ion collisions offer a unique opportunity to study strongly interacting matter under extreme conditions. The temperature and density reached in these collisions are so high that QCD becomes weakly coupled enough that quarks and gluons become deconfined and propagate as quasi-free particles. This new phase of strongly interacting matter is called the quark-gluon plasma. Most of the research in heavy ion physics at high energies revolves around the questions how the QGP comes into existence and what its properties are.

Heavy ion collisions proceed through a series of distinct phases. Important characteristics of the collision are defined in the initial state prior to the actual collision. These include the nuclear geometry, which is usually modeled with Glauber models. Soft particles production probes the gluon density at small x where gluon saturation is expected to set in. This has lead to the development of the Colour Glass Condensate as a framework for modeling heavy ion collisions in terms of saturated gluon fields. Directly after the collision of the nuclei the produced system is far from thermal equilibrium and rapidly expanding in the longitudinal direction. The early times dynamics can be described at weak coupling using kinetic theory and at strong coupling using the AdS/CFT correspondence. Both scenarios lead to equilibration of the produced matter to the extent that viscous hydrodynamics becomes applicable on the phenomenologically required (proper) time scale $\tau \lesssim 1\,\mathrm{fm/c}$. What follows is an extended phase of hydrodynamic expansion during which the system develops collective flow that is reflected in the momentum distributions of final state hadrons. The QGP has a very low shear viscosity to entropy density ratio $\eta/s \gtrsim 1/4\pi$ closed to the conjectured lower bound. The QGP is thus the least dissipative system presently known: it behaves like an almost perfect liquid. The small shear viscosity is indicative of strong residual interactions among the partons in the QGP. Finally the system has cooled so much that it transitions into the hadronic phase at the pseudo-critical temperature $T_\mathrm{c} \simeq 155\,\mathrm{MeV}$. At this point the strong coupling is so large again that there is little first principles understanding of the hadronisation process and one has to rely on phenomenological models. Hydrodynamic calculation rely in the Cooper-Frye prescription to convert the energy-momentum tensor of the fluid into hadron spectra. One idea that turned out to be useful in the context of heavy ion collisions is quark coalescence that models how quarks and anti-quarks combine to from hadrons. Scattering continues for a while in the hadronic phase and can be modeled with hydrodynamics or transport theory.

The production of jets, which are the manifestations of highly energetic partons, is found to be suppressed in heavy ion collisions — a phenomenon that is called jet quenching. The partons fragmenting into jets are produced very early in the collision in hard scattering processes that are well understood theoretically. They then propagate through the QGP and interact strongly with it. This leads to energy loss of the hard partons in what can be viewed as the partial thermalisation of the hard partons. The lost energy is transferred to the QGP and manifests itself in the form of additional soft particles situated at large angles from the jet axis. It is an open question whether the hard partons resolve quasi-particles in the QGP, i.e. whether their interaction with the QGP is best characterised by weak or strong coupling dynamics. The interpretation of jet quenching measurements is hindered by the selection bias, which biases samples of jets in heavy ion collisions towards the least modified jets and a hard fragmentation pattern.

Quarkonium states are also suppressed by colour screening in the QGP, but a quantitative understanding of the data is complicated by a number of confounding factors, most notably the regeneration of charmonium states by statistical hadronisation.

Soft particles produced in small collision systems such as high multiplicity p+p and p+A collisions show surprisingly many features such as correlations and strangeness enhancement that were thought to be signs for QGP formation. On the other hand, no jet quenching has so far been observed in small collision systems. It is not clear whether the effects come from correlations in the initial state that get imprinted on the final state or from final state interactions among particles or strings.

Ultra-peripheral collisions come with extremely strong electromagnetic fields and allow to study a large variety of processes with incoming photons. One example is the first evidence for light-by-light scattering.

Heavy ion physics is a field with a rich and diverse phenomenology that requires conceptually new approaches both theoretically and experimentally. The theoretical tools currently used include classical field theory, thermal field theory, AdS/CFT techniques, kinetic theory, relativistic hydrodynamics, effective field theories, lattice QCD and phenomenological models. This list is not complete and is likely to get longer as progress is made in understanding the many aspects of heavy ion collisions.

References

[1] J. D. Bjorken, Phys. Rev. D **27** (1983), 140-151 doi:10.1103/PhysRevD.27.140

[2] J. Adam *et al.* [ALICE], Phys. Rev. C **94** (2016) no.3, 034903 doi:10.1103/PhysRevC.94.034903 [arXiv:1603.04775 [nucl-ex]].

[3] A. Bazavov *et al.* [HotQCD], Phys. Rev. D **90** (2014), 094503 doi:10.1103/PhysRevD.90.094503 [arXiv:1407.6387 [hep-lat]].

[4] A. Aprahamian, A. Robert, H. Caines, G. Cates, J. A. Cizewski, V. Cirigliano, D. J. Dean, A. Deshpande, R. Ent and F. Fahey, *et al.*

[5] W. Busza, K. Rajagopal and W. van der Schee, Ann. Rev. Nucl. Part. Sci. **68** (2018), 339-376 doi:10.1146/annurev-nucl-101917-020852 [arXiv:1802.04801 [hep-ph]].

[6] U. W. Heinz, [arXiv:hep-ph/0407360 [hep-ph]].

[7] N. Armesto and E. Scomparin, Eur. Phys. J. Plus **131** (2016) no.3, 52 doi:10.1140/epjp/i2016-16052-4 [arXiv:1511.02151 [nucl-ex]].

[8] P. Foka and M. A. Janik, Rev. Phys. **1** (2016), 154-171 doi:10.1016/j.revip.2016.11.002 [arXiv:1702.07233 [hep-ex]].

[9] A. Andronic, P. Braun-Munzinger, K. Redlich and J. Stachel, Nature **561** (2018) no.7723, 321-330 doi:10.1038/s41586-018-0491-6 [arXiv:1710.09425 [nucl-th]].

[10] B. Abelev *et al.* [ALICE], Phys. Rev. C **88** (2013) no.4, 044909 doi:10.1103/PhysRevC.88.044909 [arXiv:1301.4361 [nucl-ex]].

[11] M. L. Miller, K. Reygers, S. J. Sanders and P. Steinberg, Ann. Rev. Nucl. Part. Sci. **57** (2007), 205-243 doi:10.1146/annurev.nucl.57.090506.123020 [arXiv:nucl-ex/0701025 [nucl-ex]].

[12] K. J. Eskola, K. Kajantie and J. Lindfors, Nucl. Phys. B **323** (1989), 37-52 doi:10.1016/0550-3213(89)90586-5

[13] E. Iancu, A. Leonidov and L. McLerran, [arXiv:hep-ph/0202270 [hep-ph]].

[14] F. Gelis, E. Iancu, J. Jalilian-Marian and R. Venugopalan, Ann. Rev. Nucl. Part. Sci. **60** (2010), 463-489 doi:10.1146/annurev.nucl.010909.083629 [arXiv:1002.0333 [hep-ph]].

[15] B. Schenke, P. Tribedy and R. Venugopalan, Phys. Rev. Lett. **108** (2012), 252301 doi:10.1103/PhysRevLett.108.252301 [arXiv:1202.6646 [nucl-th]].

[16] A. Kurkela, Nucl. Phys. A **956** (2016), 136-143 doi:10.1016/j.nuclphysa.2016.01.069 [arXiv:1601.03283 [hep-ph]].

[17] A. Kurkela and Y. Zhu, Phys. Rev. Lett. **115** (2015) no.18, 182301 doi:10.1103/PhysRevLett.115.182301 [arXiv:1506.06647 [hep-ph]].

[18] P. B. Arnold, G. D. Moore and L. G. Yaffe, JHEP **01** (2003), 030 doi:10.1088/1126-6708/2003/01/030 [arXiv:hep-ph/0209353 [hep-ph]].

[19] P. M. Chesler and W. van der Schee, Int. J. Mod. Phys. E **24** (2015) no.10, 1530011 doi:10.1142/S0218301315300118 [arXiv:1501.04952 [nucl-th]].

[20] H. Song, S. Bass and U. W. Heinz, Phys. Rev. C **89** (2014) no.3, 034919 doi:10.1103/PhysRevC.89.034919 [arXiv:1311.0157 [nucl-th]].

[21] U. W. Heinz, Landolt-Bornstein **23** (2010), 240 doi:10.1007/978-3-642-01539-7_9 [arXiv:0901.4355 [nucl-th]].

[22] P. Romatschke, Int. J. Mod. Phys. E **19** (2010), 1-53 doi:10.1142/S0218301310014613 [arXiv:0902.3663 [hep-ph]].

[23] B. Schenke, S. Jeon and C. Gale, Phys. Rev. Lett. **106** (2011), 042301 doi:10.1103/PhysRevLett.106.042301 [arXiv:1009.3244 [hep-ph]].

[24] P. Romatschke and U. Romatschke, Phys. Rev. Lett. **99** (2007), 172301 doi:10.1103/PhysRevLett.99.172301 [arXiv:0706.1522 [nucl-th]].

[25] P. Kovtun, D. T. Son and A. O. Starinets, Phys. Rev. Lett. **94** (2005), 111601 doi:10.1103/PhysRevLett.94.111601 [arXiv:hep-th/0405231 [hep-th]].

[26] K. Aamodt *et al.* [ALICE], Phys. Lett. B **708** (2012), 249-264 doi:10.1016/j.physletb.2012.01.060 [arXiv:1109.2501 [nucl-ex]].

[27] G. Nijs and W. van der Schee, SciPost Phys. **15** (2023) no.2, 041 doi:10.21468/SciPostPhys.15.2.041 [arXiv:2112.13771 [nucl-th]].

[28] F. Cooper and G. Frye, Phys. Rev. D **10** (1974), 186 doi:10.1103/PhysRevD.10.186

[29] V. Greco, C. M. Ko and P. Levai, Phys. Rev. C **68** (2003), 034904 doi:10.1103/PhysRevC.68.034904 [arXiv:nucl-th/0305024 [nucl-th]].

[30] C. B. Dover, U. W. Heinz, E. Schnedermann and J. Zimanyi, Phys. Rev. C **44** (1991), 1636-1654 doi:10.1103/PhysRevC.44.1636

[31] M. Bleicher, E. Zabrodin, C. Spieles, S. A. Bass, C. Ernst, S. Soff, L. Bravina, M. Belkacem, H. Weber and H. Stoecker, *et al.* J. Phys. G **25** (1999), 1859-1896 doi:10.1088/0954-3899/25/9/308 [arXiv:hep-ph/9909407 [hep-ph]].

[32] J. Weil *et al.* [SMASH], Phys. Rev. C **94** (2016) no.5, 054905 doi:10.1103/PhysRevC.94.054905 [arXiv:1606.06642 [nucl-th]].

[33] K. J. Eskola, P. Paakkinen, H. Paukkunen and C. A. Salgado, Eur. Phys. J. C **77** (2017) no.3, 163 doi:10.1140/epjc/s10052-017-4725-9 [arXiv:1612.05741 [hep-ph]].

[34] I. Schienbein, J. Y. Yu, K. Kovarik, C. Keppel, J. G. Morfin, F. Olness and J. F. Owens, Phys. Rev. D **80** (2009), 094004 doi:10.1103/PhysRevD.80.094004 [arXiv:0907.2357 [hep-ph]].

[35] D. de Florian, R. Sassot, P. Zurita and M. Stratmann, Phys. Rev. D **85** (2012), 074028 doi:10.1103/PhysRevD.85.074028 [arXiv:1112.6324 [hep-ph]].

[36] T. Matsui and H. Satz, Phys. Lett. B **178** (1986), 416-422 doi:10.1016/0370-2693(86)91404-8

[37] A. Tumasyan *et al.* [CMS], [arXiv:2303.17026 [hep-ex]].

[38] A. Andronic, P. Braun-Munzinger, K. Redlich and J. Stachel, Phys. Lett. B **652** (2007), 259-261 doi:10.1016/j.physletb.2007.07.036 [arXiv:nucl-th/0701079 [nucl-th]].

[39] S. Chatrchyan *et al.* [CMS], JHEP **03** (2015), 022 doi:10.1007/JHEP03(2015)022 [arXiv:1410.4825 [nucl-ex]].

[40] G. Aad *et al.* [ATLAS], Eur. Phys. J. C **79** (2019) no.11, 935 doi:10.1140/epjc/s10052-019-7439-3 [arXiv:1907.10414 [nucl-ex]].

[41] M. Aaboud *et al.* [ATLAS], Phys. Lett. B **790** (2019), 108-128 doi:10.1016/j.physletb.2018.10.076 [arXiv:1805.05635 [nucl-ex]].

[42] S. Chatrchyan *et al.* [CMS], Phys. Lett. B **712** (2012), 176-197 doi:10.1016/j.physletb.2012.04.058 [arXiv:1202.5022 [nucl-ex]].

[43] S. Chatrchyan *et al.* [CMS], Phys. Lett. B **730** (2014), 243-263 doi:10.1016/j.physletb.2014.01.042 [arXiv:1310.0878 [nucl-ex]].

[44] G. Aad *et al.* [ATLAS], Phys. Lett. B **739** (2014), 320-342 doi:10.1016/j.physletb.2014.10.065 [arXiv:1406.2979 [hep-ex]].

[45] J. Casalderrey-Solana, J. G. Milhano and U. A. Wiedemann, J. Phys. G **38** (2011), 035006 doi:10.1088/0954-3899/38/3/035006 [arXiv:1012.0745 [hep-ph]].

[46] P. M. Chesler, K. Jensen, A. Karch and L. G. Yaffe, Phys. Rev. D **79** (2009), 125015 doi:10.1103/PhysRevD.79.125015 [arXiv:0810.1985 [hep-th]].

[47] P. M. Chesler and K. Rajagopal, JHEP **05** (2016), 098 doi:10.1007/JHEP05(2016)098 [arXiv:1511.07567 [hep-th]].

[48] R. Baier, D. Schiff and B. G. Zakharov, Ann. Rev. Nucl. Part. Sci. **50** (2000), 37-69 doi:10.1146/annurev.nucl.50.1.37 [arXiv:hep-ph/0002198 [hep-ph]].

[49] J. Casalderrey-Solana, Y. Mehtar-Tani, C. A. Salgado and K. Tywoniuk, Phys. Lett. B **725** (2013), 357-360 doi:10.1016/j.physletb.2013.07.046 [arXiv:1210.7765 [hep-ph]].

[50] A. M. Sirunyan *et al.* [CMS], JHEP **05** (2018), 006 doi:10.1007/JHEP05(2018)006 [arXiv:1803.00042 [nucl-ex]].

[51] R. Kunnawalkam Elayavalli and K. C. Zapp, JHEP **07** (2017), 141 doi:10.1007/JHEP07(2017)141 [arXiv:1707.01539 [hep-ph]].

[52] C. Park, S. Jeon and C. Gale, Nucl. Phys. A **982** (2019), 643-646 doi:10.1016/j.nuclphysa.2018.10.057 [arXiv:1807.06550 [nucl-th]].

[53] Y. T. Chien and I. Vitev, JHEP **05** (2016), 023 doi:10.1007/JHEP05(2016)023 [arXiv:1509.07257 [hep-ph]].

[54] Y. Tachibana, N. B. Chang and G. Y. Qin, Phys. Rev. C **95** (2017) no.4, 044909 doi:10.1103/PhysRevC.95.044909 [arXiv:1701.07951 [nucl-th]].

[55] V. Khachatryan *et al.* [CMS], Phys. Lett. B **765** (2017), 193-220 doi:10.1016/j.physletb.2016.12.009 [arXiv:1606.06198 [nucl-ex]].

[56] J. Adam *et al.* [ALICE], Nature Phys. **13** (2017), 535-539 doi:10.1038/nphys4111 [arXiv:1606.07424 [nucl-ex]].

[57] R. D. Weller and P. Romatschke, Phys. Lett. B **774** (2017), 351-356 doi:10.1016/j.physletb.2017.09.077 [arXiv:1701.07145 [nucl-th]].

[58] A. Kurkela, U. A. Wiedemann and B. Wu, Phys. Lett. B **783** (2018), 274-279 doi:10.1016/j.physletb.2018.06.064 [arXiv:1803.02072 [hep-ph]].

[59] B. Schenke, S. Schlichting and R. Venugopalan, Phys. Lett. B **747** (2015), 76-82 doi:10.1016/j.physletb.2015.05.051 [arXiv:1502.01331 [hep-ph]].

[60] C. Bierlich, G. Gustafson and L. Lönnblad, Phys. Lett. B **779** (2018), 58-63 doi:10.1016/j.physletb.2018.01.069 [arXiv:1710.09725 [hep-ph]].

[61] L. He, T. Edmonds, Z. W. Lin, F. Liu, D. Molnar and F. Wang, Phys. Lett. B **753** (2016), 506-510 doi:10.1016/j.physletb.2015.12.051 [arXiv:1502.05572 [nucl-th]].

[62] J. Adam *et al.* [ALICE], Phys. Rev. C **91** (2015) no.6, 064905 doi:10.1103/PhysRevC.91.064905 [arXiv:1412.6828 [nucl-ex]].

[63] S. Chatrchyan *et al.* [CMS], Eur. Phys. J. C **74** (2014) no.7, 2951 doi:10.1140/epjc/s10052-014-2951-y [arXiv:1401.4433 [nucl-ex]].

[64] N. Armesto, D. C. Gülhan and J. G. Milhano, Phys. Lett. B **747** (2015), 441-445 doi:10.1016/j.physletb.2015.06.032 [arXiv:1502.02986 [hep-ph]].

[65] M. Aaboud *et al.* [ATLAS], Nature Phys. **13** (2017) no.9, 852-858 doi:10.1038/nphys4208 [arXiv:1702.01625 [hep-ex]].

[66] G. Aad *et al.* [ATLAS], JHEP **03** (2021), 243 [erratum: JHEP **11** (2021), 050] doi:10.1007/JHEP11(2021)050 [arXiv:2008.05355 [hep-ex]].

[67] A. Tumasyan *et al.* [CMS], Phys. Rev. Lett. **131** (2023), 151803 doi:10.1103/PhysRevLett.131.151803 [arXiv:2206.05192 [nucl-ex]].

[68] G. Aad *et al.* [ATLAS], Phys. Rev. Lett. **131** (2023) no.15, 151802 doi:10.1103/PhysRevLett.131.151802 [arXiv:2204.13478 [hep-ex]].

[69] J. Adam *et al.* [STAR], Phys. Rev. Lett. **127** (2021) no.5, 052302 doi:10.1103/PhysRevLett.127.052302 [arXiv:1910.12400 [nucl-ex]].

[70] I. Xu, N. Lewis, X. Wang, J. D. Brandenburg and L. Ruan, [arXiv:2211.02132 [hep-ex]].

[71] S. Acharya *et al.* [ALICE], Phys. Lett. B **817** (2021), 136280 doi:10.1016/j.physletb.2021.136280 [arXiv:2101.04623 [nucl-ex]].

[72] S. Acharya *et al.* [ALICE], [arXiv:2305.06169 [nucl-ex]].

Scientific programme[1]

Field Theory & the Electro-Weak Standard Model

Anna Kulesza (University of Münster, Germany)

QCD

Gudrun Heinrich (KIT, Germany)

Cosmology

Mikhail Shaposhnikov (EPFL, Switzerland)

Practical Statistics

Troels C. Petersen (Niels Bohr Institute, Denmark)

Special Lecture on Dark Matter

Mads Frandsen (SDU Galaxy, Denmark)

Higgs and Beyond

Christophe Grojean (DESY and Humboldt University of Berlin, Germany)

Neutrino Physics

Gabriela Barenboim (University of Valencia, Spain)

Flavour Physics and CP Violation

Alexander Lenz (University of Siegen, Germany)

Machine Learning for HEP

Jan Kieseler (KIT, Germany)

Heavy-Ion Physics

Korinna Zapp (Lund University, Sweden)

Prospects at LHC in Run 3 and HL-LHC

Pamela Ferrari (Nikhef, The Netherlands)

Question & Answer Session with CERN Director-General

Fabiola Gianotti (CERN)

Outreach Training

(Inside Edge)

[1] Slides available at https://indico.cern.ch/event/1256499/.

Organizing committees

Standing committee

Nick Ellis (CERN)
Markus Elsing (CERN)
Alexander Huss (CERN)
Martijn Mulders (CERN)
Kate Ross (CERN)
Sascha Stahl (CERN)

Local organizing committee

Ulrik Ingerslev Uggerhøj (Aarhus University)
Hans Otto Uldall Fynbo (Aarhus University)
Hanne Bak (Aarhus University)
Louise Kindt (University of Southern Denmark)
Mads Toudal Frandsen (University of Southern Denmark)
Mikkel Theiss Kristensen (University of Southern Denmark)
John Renner Hansen (Copenhagen University)
Malene Emilie Maar Vinding (Copenhagen University)
Peter Hansen (Copenhagen University)
Jens Jørgen Gaardhøje (Copenhagen University)
Stefania Xella (Copenhagen University)

List of lecturers

Gabriela Barenboim (University of Valencia and IFIC, Spain)
Pamela Ferrari (Nikhef, The Netherlands)
Mads Frandsen (SDU Galaxy, Denmark)
Fabiola Gianotti (CERN)
Christophe Grojean (DESY and Humboldt University of Berlin, Germany)
Gudrun Heinrich (KIT, Germany)
Inside Edge
Jan Kieseler (KIT, Germany)
Anna Kulesza (University of Münster, Germany)
Alexander Lenz (University of Siegen, Germany)
Troels C. Petersen (Niels Bohr Institute, Denmark)
Mikhail Shaposhnikov (EPFL, Switzerland)
Korinna Zapp (Lund University, Sweden)

List of discussion leaders

Emanuele Bagnaschi (CERN)
Tyler Corbett (Vienna, Austria)
Tomasz Dutka (KIAS, South Korea)
Maximilian Loeschner (DESY, Germany)
Alba Soto Ontoso (CERN)
Giovanni Stagnitto (Zurich, Switzerland)

List of Students

Erlend Aakvaag

Anke Ackermann

Marijus Ambrozas

Maura Barros

Chiara Basile

Daariimaa Battulga

Gaya Benane

Hicham Benmansour

Tiziano Bevilacqua

Naman Kumar Bhalla

Kartik Deepak Bhide

Shyam Bhuller

Benjamin Blancon

Léo Boudet

Aodhan Burke

Antimo Cagnotta

Long Hoa Cao Phuc

Alberto Carnelli

Michael William Carrigan

Daniele Centanni

Jieun Choi

Artur Cordeiro Oudot Choi

Valeria D'Amante

Soumya Dansana

Christina Dimitriadi

Ahmed Eddymaoui

Levi Evans

Federica Fabiano

Luis Falda Coelho

Arianna Gemma Garcia Caffaro

Svetlana Gertsenberger

Gediminas Glemza

Benedikt Gocke

Mateusz Goncerz

Matteo Greco

Bennett Greenberg

Christopher Greenberg

Simon Gabriel Grewe

Giovanni Guerrieri

Lei Hao

Paula Herrero Gascón

Xiaonan Hou

Tao Hsu

Hendrik Jage

Prasham Jain

Fiona Ann Jolly

Gaelle Khreich

Seulgi Kim

Timothy Michael Knight

Finn Jonathan Labe

Roxani Lazaridou

Gauthier Legras

Dennis Lindebaum

Ryan McCarthy

Niall Thomas McHugh

Marin Mlinarevic

Attia Mohamed

Federico Montereali

Arnau Morancho Tarda

Yasaman Najafijozani

Keerthi Nakkalil

Sahana Narasimha

Małgorzata Niemiec

Anja Novosel

Ivan Pidhurskyi

Andres Pinto

Elena Pompa Pacchi

Andris Potrebko

Nathan Prouvost

Ian Reed

Selaiman Ridouani

Graziella Russo

Maria Adriana Sabia

Valentina Sarkisovi

Ana Sculac

Konstantin Sharko

Zhihong Shen

Boan Shi

Supriya Sinha

Cristiano Tarricone

Andrea Tavira Garcia

Wesley Terrill

Mattias Ermakov Thing

Kevin Urquía

Katharina Voss

J Alexander Ward

Harriet Watson

Aidan Richard Wiederhold

Lukas Witola

Yoran Yeh

Saeahram Yoo

Zhenxuan Zhang

André Zimermmane-Santos

List of Posters

Poster title	**Presenter**
Search for dark matter produced in association with a Higgs boson decaying to tau leptons at $\sqrt{s} = 13$ TeV with the ATLAS detector	ERLEND AAKVAAG
Measurement of the $ZZ\gamma$ final state with the ATLAS detector at the LHC	ANKE ACKERMANN
ATLAS Forward Proton Detector as a Tool for New Physics Searches	MAURA BARROS
Search for $H(\gamma\gamma) + c$ production at CMS	TIZIANO BEVILACQUA
Measuring The Luminosity Of Heavy Ion Collisions With A New Algorithm Using Charged Particle Tracks At The ATLAS Detector	KARTIK DEEPAK BHIDE
Signature-based search for new physics producing a top quark plus invisible particles using ML techniques at the CMS experiment	ANTIMO CAGNOTTA
Measurement differential cross sections of $Z\gamma$ in proton-proton collisions at $s = \sqrt{13}$TeV	LONG HOA CAO PHUC
SND@LHC: The Scattering and Neutrino Detector at the LHC	DANIELE CENTANNI
Searching for long lived axion like particles through Higgs boson decays using ATLAS calorimeters	ARTUR CORDEIRO OUDOT CHOI
Search for resonant di-Higgs production in the $bb\tau\tau$ final state at CMS and tau identification at the CMS HLT	VALERIA D'AMANTE
A Top Friendship: Searching for $t\bar{t}H(\rightarrow b\bar{b})$ with ATLAS	LEVI EVANS
Fixed-target collisions at LHCb	FEDERICA FABIANO
Phase-II Tracking for ITk with ACTS and Seach for $ttHH(HH \rightarrow 4b)$ Production	LUIS FALDA COELHO
Perturbatively Regularized Neural Networks	ARIANNA GEMMA GARCIA CAFFARO
Lepton flavor violation study in the NA64 experiment	SVETLANA GERTSENBERGER
Measurement of top quark involved CKM matrix elements in single top-quark t-channel processes	BENEDIKT GOCKE
Searches for supersymmetric particles at the ATLAS detector and misalignment studies of the New Small Wheel	MATTEO GRECO

Poster title	**Presenter**
Data Scouting and Parking at the CMS High-Level Trigger	BENNETT PAUL GREENBERG
Search For Charged Higgs Boson In $H^{\pm} \to W^{\pm}h$ Decays With The ATLAS Detector	SIMON GABRIEL GREWE
Measurement of CKM angle γ using $B^{\pm} \to D^{*}h^{\pm}$ decays	LEI HAO
Charm production asymmetries in the LHCb experiment	PAULA HERRERO GASCON
New helium identification technique at LHCb	HENDRIK JAGE
Study of polarized same-sign WW production with the ATLAS detector	PRASHAM JAIN
Angular Analysis of the $B_s^0 \to \phi(K^+K^-)e^+e^-$ decay at LHCb	GAELLE KHREICH
Development of ML-based topological algorithms for the CMS level-1 trigger	FINN JONATHAN LABE
Higgs boson production properties via the $H \to \tau\tau$ channel	ROXANI LAZARIDOU
Displaced Vertex Track Trigger for the CMS Phase-2 Upgrade	RYAN EDWARD MCCARTHY
Search for time-dependent CP violation in $D^0 \to \pi^+\pi^-\pi^0$ decays at LHCb	NIALL THOMAS MCHUGH
Search for nonresonant pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state with ATLAS	MARIN MLINAREVIC
Study of the Kinematic Fit algorithm for the $HH \to b\bar{b}\gamma\gamma$ channel	FEDERICO MONTEREALI
Deriving a ML-based (b-)jet energy regression using data driven techniques	ARNAU MORANCHO TARDA
SModelS: enabling global likelihood analyses	SAHANA NARASIMHA
Transverse Momentum Dependent Transverse Spin Asymmetries in COMPASS Drell-Yan data	MAŁGORZATA NIEMIEC
Prompt Lepton Jets at ATLAS Run2	ELENA POMPA PACCHI
Flavor-dependent (L5) MC truth jet energy corrections and flavor uncertainties in CMS Run 2	ANDRIS POTREBKO
The High-Granularity Timing Detector for ATLAS Phase-II upgrade	SELAIMAN RIDOUANI
Implementation of low pt electron reconstruction in $HZZ4l$ analysis	ANA SCULAC

Poster title	**Presenter**
Precision measurement of the $t\bar{t}$ production cross section	Konstantin Sharko
Search for hidden-charm tetraquark candidates with strangeness at LHCb	Zhihong Shen
Helicity amplitude analysis of $\chi_{cJ} \to \phi\phi$	Boan Shi
A sensitivity study of triboson production to dimension-six EFT operators at the LHC	Cristiano Tarricone
Studying Double Charm Production with ALICE	Andrea Tavira Garcia
A Search for New Physics in Final States with Two or Three Soft Leptons and Missing p_T	Wesley Thomas Terrill
Dark matter models via Higgs portals	Mattias Ermakov Thing
Towards a modelling systematic uncertainty recipe for $bb4\ell$	Katharina Voss
Angular analysis of $B^0_{(s)} \to \pi^+\pi^-\mu^+\mu^-$ decays at LHCb	J Alexander Ward
Differential cross-section measurements of $t\bar{t}Z$ with the ATLAS detector	Harriet Watson
Measuring the CKM Angle γ Using $B^0 \to DK^+\pi^-$ Decays at LHCb	Aidan Richard Wiederhold
The LHCb Scintillating Fibre Tracker	Lukas Witola
Contur: Constraints on new theories using Rivet	Yoran Yeh
Status of Υ Polarization Studies in PbPb Collisions with CMS	Saeahram Yoo