CERN–2014–001
KEK Proceedings 2013–8
17 March 2014

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

2012 Asia–Europe–Pacific School of High-Energy Physics

Fukuoka, Japan
14 – 27 October 2012

Proceedings

Editors: K. Kawagoe
M. Mulders

GENEVA
2014

ISBN 978–92–9083–399–4
ISSN 0531–4283

This report should be cited as:
Proceedings of the 2012 Asia–Europe–Pacific School of High-Energy Physics, Fukuoka, Japan, 14–27 October 2012, edited by K. Kawagoe and M. Mulders, CERN–2014–001 and KEK-Proceedings 2013–8 (CERN, Geneva, 2014), DOI:10.5170/CERN-2014-001

A contribution in this report should be cited as:
[Author name(s)], in Proceedings of the 2012 Asia–Europe–Pacific School of High-Energy Physics, Fukuoka, Japan, 14–27 October 2012 , edited by K. Kawagoe and M. Mulders, CERN–2014–001 and KEK-Proceedings 2013–8, (CERN, Geneva, 2014), pp. [first page]–[last page], DOI:10.5170/CERN-2014-001.[first page]

Abstract

The Asia–Europe–Pacific School of High-Energy Physics is intended to give young physicists an introduction to the theoretical aspects of recent advances in elementary particle physics. These proceedings contain lectures on quantum field theory, quantum chromodynamics, flavour physics and CP-violation, physics beyond the Standard Model, neutrino physics, particle cosmology, heavy-ion physics, as well as a presentation of recent results from the Large Hadron Collider (LHC), practical statistics for particle physicists and a short introduction to the principles of particle physics instrumentation.

Preface

The first event in the new series of the Asia–Europe–Pacific School of High-Energy Physics took place in Fukuoka, Japan, from 14 to 27 October 2012. A strong team from KEK, as well as from Kyushu and Saga Universities, provided excellent local organization. CERN and KEK collaborated to provide administrative support in preparation for the School.

The staff and students were housed in comfortable accommodation in the Luigans spa and resort complex that also provided excellent conference facilities. The students shared twin or three-bed rooms, mixing nationalities to foster cultural exchange between participants from different countries.

A total of 83 students coming from 21 different countries attended the school. About 70% of the students were from Asia-Pacific countries, most of the others coming from Europe. More than 80% of the participants were working towards a PhD, while most of the others were advanced Masters students; the School was also open to postdocs. Over 80% of the students were experimentalists; the school was also open to phenomenologists.

A total of 34 lectures were complemented by daily discussion sessions led by six discussion leaders. The teachers (lecturers and discussion leaders) came from many different countries: Australia, China, France, Germany, India, Japan, Korea, Russia, Spain, Switzerland, Taiwan and the United Kingdom.

The programme required the active participation of the students. In addition to the discussion sessions that addressed questions from the lecture courses, there was an evening session in which many students presented posters about their own research work to their colleagues and the teaching staff. The high level of interest could be gauged by the fact that discussion of the posters continued into the early hours of the next morning.

Collaborative projects in which the students of each Discussion Group worked together on an in-depth study of a published experimental data analysis were an important activity. This required interacting, outside of the formal teaching sessions, with colleagues from different countries and different cultures. A student representative of each of the six groups presented a short summary of the conclusions of the group's work in a special evening session whose attendees included the Directors General of CERN and KEK, both of whom then delivered lectures on the final day of the School.

Leisure activities included a full-day excursion to the Mount Aso volcano, and a half-day excursion to Dazaifu where the group visited a Buddhist temple and a Shinto shrine as well as the Kyushu national museum. There was also a free afternoon during which participants could visit the city of Fukuoka at the time of the Hakata Okunchi Festival.

Our thanks go to the local-organization team and, in particular, to Professor Kiyotomo Kawagoe for all his work and assistance in preparing the School, on both scientific and practical matters, and for his 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, undoubtedly appreciated listening to and discussing with the teaching staff of world renown.

We would like to express our strong appreciation to Professor Rolf Heuer, Director General of CERN, and Professor Atsuto Suzuki, Director General of KEK, for their lectures on the particle-physics programmes in Europe and in Asia, and for discussing with the School participants.

We would particularly like to thank Hiroshi Ogawa, Governor of Fukuoka Prefecture, and Professor Setsuo Arikawa, President of Kyushu University, for visiting the School and more generally for their interest and support.

We would also like to recognize the important work done by the International Advisory Committee under

the chairmanship of Professor Fumihiko Takasaki in initiating this new series of Schools, and in establishing the framework and strategy for their organization. In both the International Advisory Committee and the International Organizing Committee, there is broad representation from Asia–Pacific countries with members from Australia, China, India, Japan, Korea, Russia, Taiwan, as well as from Europe with members from CERN and France. The collaborative spirit within and between these two committees has been very important for achieving this first and successful School.

We are very grateful to Hélène Haller and Kate Ross from CERN, and Misa Miyai and Masami Yokoyama from KEK, for their untiring efforts on administration for the School.

Sponsorship from numerous bodies in many countries covered the cost of travel and/or local expenses of their staff and students who attended the School. In addition, general sponsorship is gratefully acknowledged from: CEA/Irfu, France; CNRS/IN2P3, France; CERN; Fukuoka Convention and Visitors Bureau, Japan; Fukuoka Prefecture, Japan; KEK, Japan; Kyushu University, Japan.

Nick Ellis
(On behalf of the Organizing Committee)

Representatives of the organizing committees and the Governor of Fukuoka prefecture at the opening ceremony of the School. From left to right: Professor Kiyotomo Kawagoe, Professor Fumihiko Takasaki, Governor Hiroshi Ogawa and Dr. Nick Ellis.


The 1st Asia-Europe-Pacific School of High-Energy Physics
from 14 to 27 October 2012 The Luigans Spa & Resort, Fukuoka, Japan


The 1st Asia-Europe-Pacific School of High-Energy Physics
from 14 to 27 October 2012 The Luigans Spa & Resort, Fukuoka, Japan

People in the photograph

1 Peter WIJRATNE
2 Jason LEE
3 Jeremy NEVEU
4 Morten JOERGENSEN
5 Michael PRIM
6 Simon HEISTERKAMP
7 Nikhul PATEL
8 Kin-ya ODA
9 Raymond VOLKAS
10 Takuya MINAKUCHI
11 Petr KATRENKO
12 Yuji SUDO
13 Jia Jian TEOH
14 Tsunayuki MATSUBARA
15 Shoaib KHALID
16 Kong Guan TAN
17 Kenji KIUCHI
18 Myeonghun PARK
19 Varchaswi KASHYAP
20 Benda XU
21 Lan-Chun LV
22 Kou OISHI
23 David JENNENS
24 Jinsu KIM
25 Metteo FRANCHINI
26 Yuhei ITO
28 Kenzo SUZUKI
29 Guang ZHAO
30 Sebastian WANDERNOTH
31 Elisabeth PENZENBOECK
32 Camille COUTURIER
33 Po-Yen TSENG
34 Saranya GHOSH
35 Shoichiro NISHIMURA
36 Geon-Bo KIM
37 Jongkuk KIM
38 Glang ZHAO
40 Manoj Kumar SINGH
41 Sabyasachi CHAKRABORTY
42 Zhong-Zhi XIANYU
43 Maryam HASHEMINIA
44 Jun WAKABAYASHI
45 Almut PINGEL
46 Katsuya YAMAUCHI
47 Vitaly VOROBYEV
48 Ka Hei Martin KWOK
49 Emyr CLEMENT
50 Annika VAUTH
51 Wajid KHAN
52 Cenk TÜRKOGLU
53 Shoichiro NISHIMURA
54 Harvey MADDOCS
55 Lemuel PELAGIO
56 Satoru MATSUMOTO
57 Philippe GROS
58 Thomas WILLIAMS
59 Tomoe KISHIMOTO
60 Daisuke YAMATO
61 Saurabh SANDILYA
62 Camila RANGEL SMITH
63 Su-Yin WANG
64 Kanishka RAWAT
65 Nam TRAN
66 Bei-Zhen HU
67 Hideyuki OIDE
68 Lili JIANG
70 Tomoko IWASHITA
71 Diego SEMMLER
72 Nadeesha WICKRAMAGE
73 Francesca DORDEI
74 Ali ZAMAN
75 Xiaokang ZHOU
76 Francesco RIVA
77 Elida ISTIQOMAH
78 Nooraihan ABDULLAH
79 Curtis BLACK
80 Rintarn SAENGSAI
81 Kyungwon KIM
82 Marc BESAN CON
83 Luis ALVAREZ-GAUME
84 Werner RIEGLER
85 Nick ELLIS
86 Kiyotomo KAWAGOE
87 Martijn MULDERS
88 Jose OCARIZ
89 Lydia ROOS
90 Mehar Ali SHAN
91 Masami YOKOYAMA
92 Ian WATSON
93 Shigeki HIROSE

PHOTOGRAPHS (MONTAGE 1)

PHOTOGRAPHS (MONTAGE 2)

PHOTOGRAPHS (MONTAGE 3)

Contents

Introductory Lectures on Quantum Field Theory

Luis Álvarez-Gaumé [a] *and Miguel A. Vázquez-Mozo* [b]
[a] CERN, Geneva, Switzerland
[b] Universidad de Salamanca, Salamanca, Spain

Abstract
In these lectures we present a few topics in Quantum Field Theory in detail. Some of them are conceptual and some more practical. They have been selected because they appear frequently in current applications to Particle Physics and String Theory.

1 Introduction

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

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

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

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

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

1.1 A note about notation

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

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

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

$$\sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2}$$

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

2 Why do we need Quantum Field Theory after all?

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

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

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

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

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

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

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

$$j_\mu = \frac{i}{2}\Big(\psi^*\partial_\mu\psi - \partial_\mu\psi^*\,\psi\Big), \tag{6}$$

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

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

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

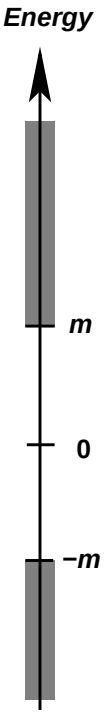


Fig. 1: Spectrum of the Klein-Gordon wave equation.

Given the inner product

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

the states (7) form an orthonormal basis

$$\begin{aligned}\langle f_p|f_{p'}\rangle &= \delta(\vec{p}-\vec{p}\,'),\\ \langle f_{-p}|f_{-p'}\rangle &= -\delta(\vec{p}-\vec{p}\,'), \qquad (8)\\ \langle f_p|f_{-p'}\rangle &= 0. \qquad (9)\end{aligned}$$

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

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

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

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

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

where $\vec{\alpha}$ and β are 4×4 matrices

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

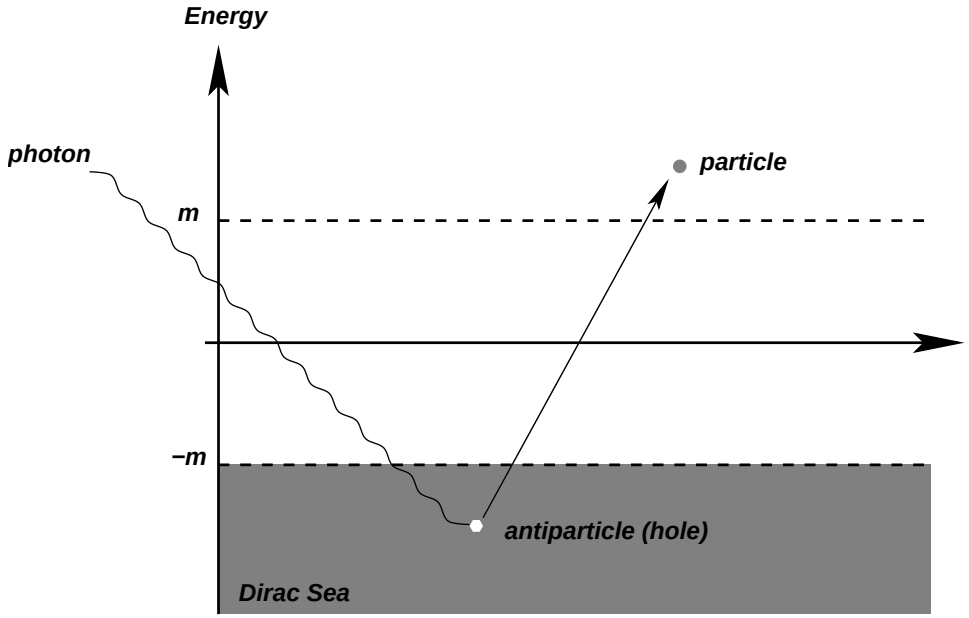


Fig. 2: Creation of a particle-antiparticle pair in the Dirac sea picture.

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

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

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

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

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

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

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

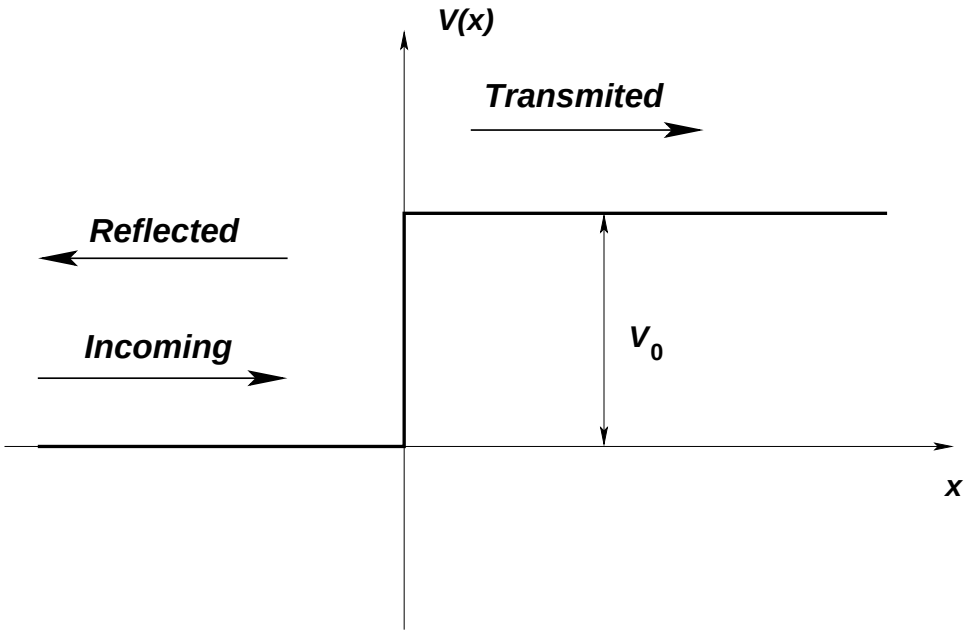


Fig. 3: Illustration of the Klein paradox.

where the mass-shell condition implies that

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

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

$$T = \frac{2p_1}{p_1 + p_2}, \qquad R = \frac{p_1 - p_2}{p_1 + p_2}. \tag{16}$$

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

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

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

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

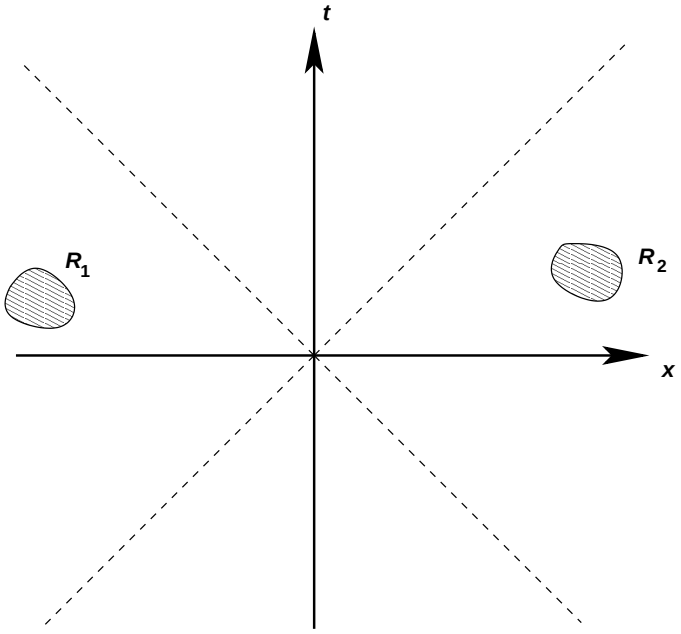


Fig. 4: Two regions R_1, R_2 that are causally disconnected.

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

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

$$\Delta E \sim m \tag{17}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3 From classical to quantum fields

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

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

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

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

$$\int d^3p \, |\vec{p}\rangle \langle \vec{p}| = \mathbf{1} \tag{26}$$

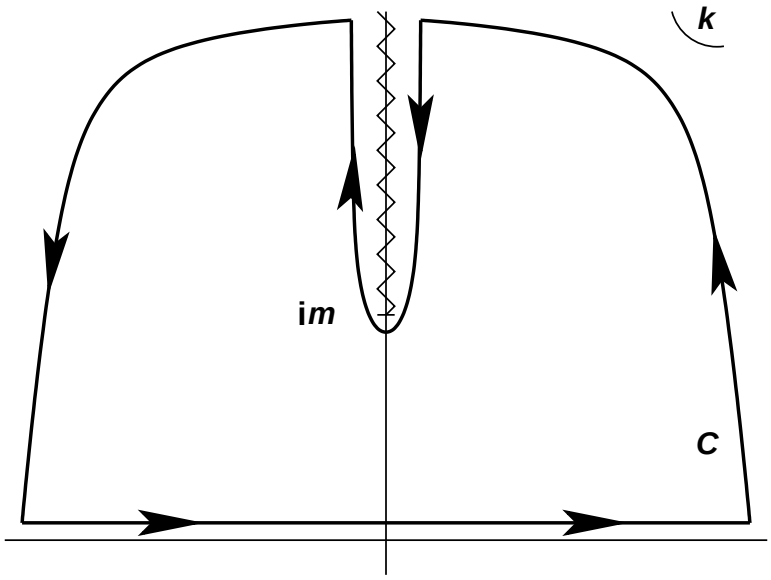


Fig. 5: Complex contour C for the computation of the integral in Eq. (24).

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

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

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

$$\widehat{P}^i = \int d^3p\, |\vec{p}\rangle\, p^i\, \langle\vec{p}| \tag{28}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

which in our case implies that

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

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

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

Hence, the relativistic invariant measure is given by

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

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

$$\widehat{P}^0|p\rangle = \omega_p|p\rangle, \qquad\qquad \widehat{P}^i|p\rangle = p^{\,i}|p\rangle. \tag{37}$$

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

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

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

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

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

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

[2]The factors of 2π are introduced for later convenience.

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

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

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

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

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

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

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

$$p^0 = m\cosh\lambda, \qquad\qquad p^1 = m\sinh\lambda. \tag{44}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

with the Lorentz invariant commutation relations

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

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

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

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

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

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

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

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

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

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

and the corresponding one for creation operators.

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

- **Hermiticity.**

$$\phi^\dagger(x) = \phi(x). \tag{56}$$

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

$$[\phi(x),\phi(y)] = 0, \qquad (x-y)^2 < 0. \tag{57}$$

- **Translation invariance.**

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

- **Lorentz invariance.**

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

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

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

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

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

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

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

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

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

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

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

where $\varepsilon(x)$ is defined as

$$\varepsilon(x) \equiv \theta(x)-\theta(-x) = \left\{ \begin{array}{cc} 1 & x>0 \\ -1 & x<0 \end{array} \right. . \tag{65}$$

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

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

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

3.1 Canonical quantization

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

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

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

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

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

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

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

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

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

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

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

Their Poisson bracket is defined by

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

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

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

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

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

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

$$i\{\cdot,\cdot\} \longrightarrow [\cdot,\cdot]. \tag{75}$$

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

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

whose general solution can be written as[3]

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

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

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

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

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

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

$$\alpha(\vec{p}) \longrightarrow \widehat{\alpha}(\vec{p}), \qquad\qquad \alpha(\vec{p})^* \longrightarrow \widehat{\alpha}^\dagger(\vec{p}). \tag{79}$$

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

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

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

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

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

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

$$E_{\text{vac}} \equiv \langle 0|\widehat{H}|0\rangle = \sum_{\vec{p}}\frac{1}{2}\omega_p. \tag{82}$$

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

3.2 The Casimir effect

The presence of a vacuum energy is not characteristic of the scalar field. It is also present in other cases, in particular in quantum electrodynamics. Although one might be tempted to discarding this infinite contribution to the energy of the vacuum as unphysical, it has observable consequences. In 1948 Hendrik

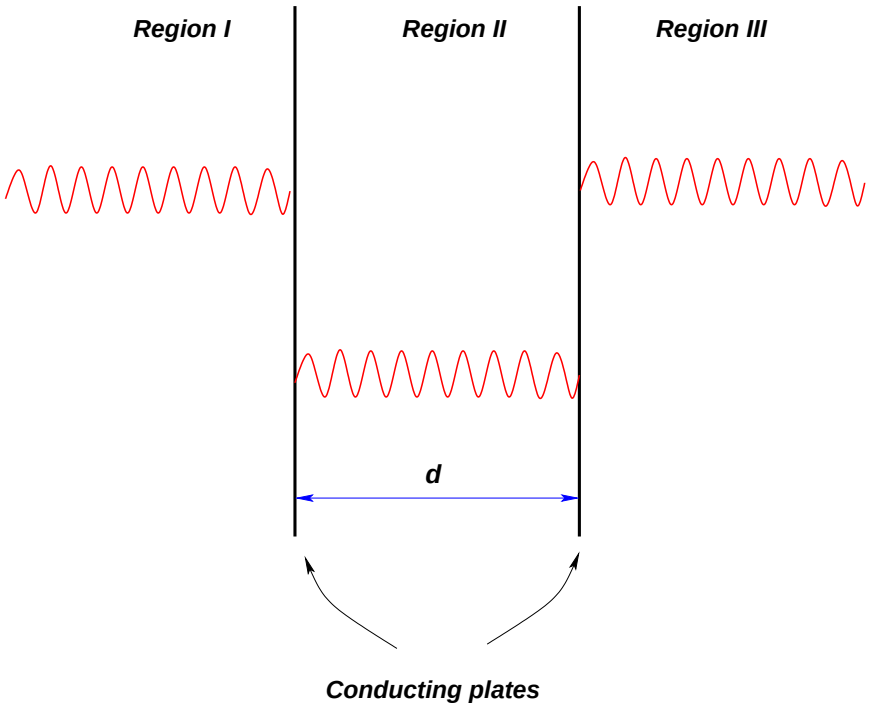


Fig. 6: Illustration of the Casimir effect. In regions I and II the spetrum of modes of the momentum $p_\perp$ is continuous, while in the space between the plates (region II) it is quantized in units of $\frac{\pi}{d}$.

Casimir pointed out [16] that although a formally divergent vacuum energy would not be observable, any variation in this energy would be (see [17] for comprehensive reviews).

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

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

$$E(d)_{\rm reg} = E(d)_{\rm vac} - E(\infty)_{\rm vac} \tag{83}$$

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

$$E(d)_{\rm reg} \;=\; S\int \frac{d^2p_\parallel}{(2\pi)^2}\frac{1}{2}|\vec{p}_\parallel| + 2S\int \frac{d^2p_\parallel}{(2\pi)^2}\sum_{n=1}^{\infty}\frac{1}{2}\sqrt{\vec{p}_\parallel^{\,2} + \left(\frac{n\pi}{d}\right)^2}$$

$$- \quad 2Sd \int \frac{d^3p}{(2\pi)^3} \frac{1}{2} |\vec{p}| \tag{84}$$

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

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

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

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

the regularized vacuum energy can be written as

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

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

$$\begin{aligned} \sum_{n=1}^{\infty} F(n) - \int_0^\infty dx\, F(x) &= -\frac{1}{2} \left[F(0) + F(\infty) \right] + \frac{1}{12} \left[F'(\infty) - F'(0) \right] \\ &- \frac{1}{720} \left[F'''(\infty) - F'''(0) \right] + \dots \end{aligned} \tag{88}$$

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

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

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

$$P_{\rm Casimir} = -\frac{\pi^2}{240} \frac{1}{d^4}. \tag{90}$$

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

4 Theories and Lagrangians

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

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

4.1 Representations of the Lorentz group

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

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

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

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

These six generators satisfy the algebra

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

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

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

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

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

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

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

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

The next interesting cases are $(\frac{\mathbf{1}}{\mathbf{2}}, \mathbf{0})$ and $(\mathbf{0}, \frac{\mathbf{1}}{\mathbf{2}})$. They correspond respectively to a right-handed and a left-handed Weyl spinor. Their properties will be studied in more detail below. In the case of $(\frac{\mathbf{1}}{\mathbf{2}}, \frac{\mathbf{1}}{\mathbf{2}})$, since from Eq. (94) we see that $J_i = J_i^+ + J_i^-$ the rules of addition of angular momentum tell us that there are two states, one of them transforming as a vector and another one as a scalar under three-dimensional rotations. Actually, a more detailed analysis shows that the singlet state corresponds to the time component of a vector and the states combine to form a vector under the Lorentz group.

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

Table 1: Representations of the Lorentz group.

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

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

$$P : J_i \longrightarrow J_i, \qquad P : M_i \longrightarrow -M_i \tag{96}$$

this means that $P : J_i^{\pm} \longrightarrow J_i^{\mp}$ and therefore a representation $(\mathbf{s_1},\mathbf{s_2})$ is transformed into $(\mathbf{s_2},\mathbf{s_1})$. This means that, for example, a vector $(\frac{\mathbf{1}}{\mathbf{2}},\frac{\mathbf{1}}{\mathbf{2}})$ is invariant under parity, whereas a left-handed Weyl spinor $(\frac{\mathbf{1}}{\mathbf{2}},\mathbf{0})$ transforms into a right-handed one $(\mathbf{0},\frac{\mathbf{1}}{\mathbf{2}})$ and vice versa.

4.2 Spinors

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

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

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

$$u_+^\dagger \sigma_+^\mu u_+, \qquad u_-^\dagger \sigma_-^\mu u_-. \tag{98}$$

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

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

$$u_\pm \longrightarrow e^{i\theta} u_\pm \tag{99}$$

we are left with just one possibility up to a sign

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

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

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

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

$$\partial_\mu \partial^\mu u_\pm = 0, \tag{102}$$

whose solutions are:

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

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

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

which implies

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

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

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

$$(\mathbf{\tfrac{1}{2}}, \mathbf{\tfrac{1}{2}}) = (\mathbf{\tfrac{1}{2}}, \mathbf{0}) \otimes (\mathbf{0}, \mathbf{\tfrac{1}{2}}) \quad \text{and} \quad (\tfrac{1}{2}, \tfrac{1}{2}) = (\mathbf{0}, \mathbf{\tfrac{1}{2}}) \otimes (\mathbf{\tfrac{1}{2}}, \mathbf{0}). \tag{106}$$

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

$$(\mathbf{\tfrac{1}{2}}, \mathbf{0}) \otimes (\mathbf{\tfrac{1}{2}}, \mathbf{0}) = (\mathbf{1}, \mathbf{0}) \oplus (\mathbf{0}, \mathbf{0}), \qquad (\mathbf{0}, \mathbf{\tfrac{1}{2}}) \otimes (\mathbf{0}, \mathbf{\tfrac{1}{2}}) = (\mathbf{0}, \mathbf{1}) \oplus (\mathbf{0}, \mathbf{0}). \tag{107}$$

The singlet representations corresponds to the antisymmetric combinations

$$\epsilon_{ab} u_\pm^a u_\pm^b, \tag{108}$$

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

At first sight it might seem that the term (108) vanishes identically because of the antisymmetry of the ϵ-symbol. However we should keep in mind that the spin-statistic theorem (more on this later) demands that fields with half-integer spin have to satisfy the Fermi-Dirac statistics and therefore satisfy anticommutation relations, whereas fields of integer spin follow the statistic of Bose-Einstein and, as a

consequence, quantization replaces Poisson brackets by commutators. This implies that the components of the Weyl fermions $u_\pm$ are anticommuting Grassmann fields

$$u_\pm^a u_\pm^b + u_\pm^b u_\pm^a = 0. \tag{109}$$

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

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

$$\mathcal{L}^\pm_{\rm Weyl} = i u_\pm^\dagger \sigma_\pm^\mu \partial_\mu u_\pm - \frac{m}{2}\epsilon_{ab} u_\pm^a u_\pm^b + \text{h.c.} \tag{110}$$

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

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

$$P : u_\pm \longrightarrow u_\mp. \tag{111}$$

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

$$\psi = \begin{pmatrix} u_+ \\ u_- \end{pmatrix} \tag{112}$$

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

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

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

These equations of motion can be derived from the Lagrangian density

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

To simplify the notation it is useful to define the Dirac γ-matrices as

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

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

$$\overline{\psi} \equiv \psi^\dagger \gamma^0 = \psi^\dagger \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}. \tag{116}$$

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

$$\mathcal{L}_{\rm Dirac} = \overline{\psi}\,(i\gamma^\mu\partial_\mu - m)\,\psi. \tag{117}$$

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

$$\gamma^0 = \beta, \qquad \gamma^i = i\alpha^i. \tag{118}$$

In addition, the γ-matrices defined in (115) satisfy the Clifford algebra

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}. \tag{119}$$

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

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

We see that γ^5 anticommutes with all other γ-matrices. This implies that

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

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

$$P_+\psi = \begin{pmatrix} u_+ \\ 0 \end{pmatrix}, \qquad P_-\psi = \begin{pmatrix} 0 \\ u_- \end{pmatrix}, \tag{122}$$

which are precisely the Weyl spinors introduced before.

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

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

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

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

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

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

In the same way, for negative energy solutions we have

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

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

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

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

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

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

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

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

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

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

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

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

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

The Hamiltonian operator for the Dirac field is

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

At this point we realize again of the necessity of quantizing the theory using anticommutators instead of commutators. Had we use canonical commutation relations, the second term inside the integral in (132) would give the number operator $d^\dagger(\vec{k},s)d(\vec{k},s)$ with a minus sign in front. As a consequence the Hamiltonian would be unbounded from below and we would be facing again the instability of the theory

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

already noticed in the context of relativistic quantum mechanics. However, because of the *anticommutation* relations (131), the Hamiltonian (132) takes the form

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

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

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

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

$$j^\mu = \overline{\psi}\gamma^\mu\psi, \qquad \partial_\mu j^\mu = 0. \tag{135}$$

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

$$Q = e\int d^3x\, j^0 \tag{136}$$

is identified with the electric charge.

4.3 Gauge fields

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

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

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

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

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

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

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

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

$$A_\mu \longrightarrow A_\mu + \partial_\mu \epsilon. \tag{140}$$

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

$$m\frac{du^\mu}{d\tau} = eF^{\mu\nu}u_\nu, \tag{141}$$

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

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

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

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

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

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

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

where Γ_1 and Γ_2 are two curves surrounding the solenoid from different sides, and Γ is any closed loop surrounding it. Therefore the relative phase between the two beams gets an extra term depending on the value of the vector potential outside the solenoid as

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

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

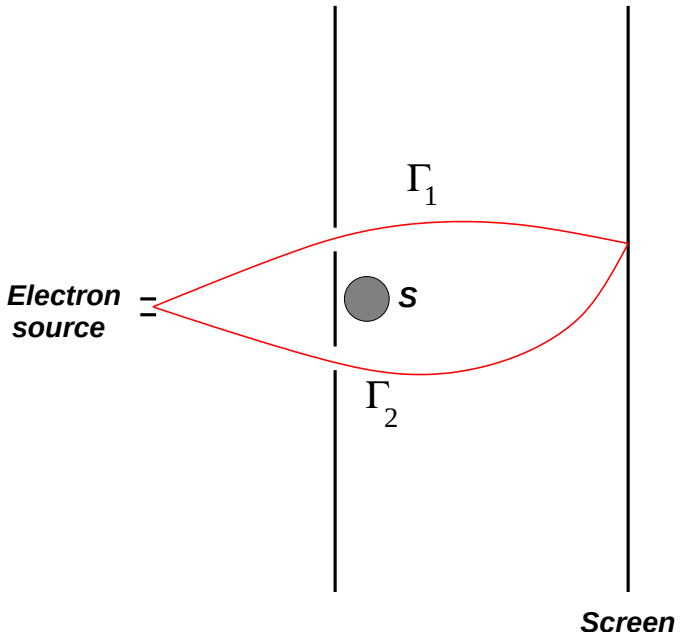


Fig. 7: Illustration of an interference experiment to show the Aharonov-Bohm effect. S represent the solenoid in whose interior the magnetic field is confined.

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

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

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

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

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

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

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

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

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

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

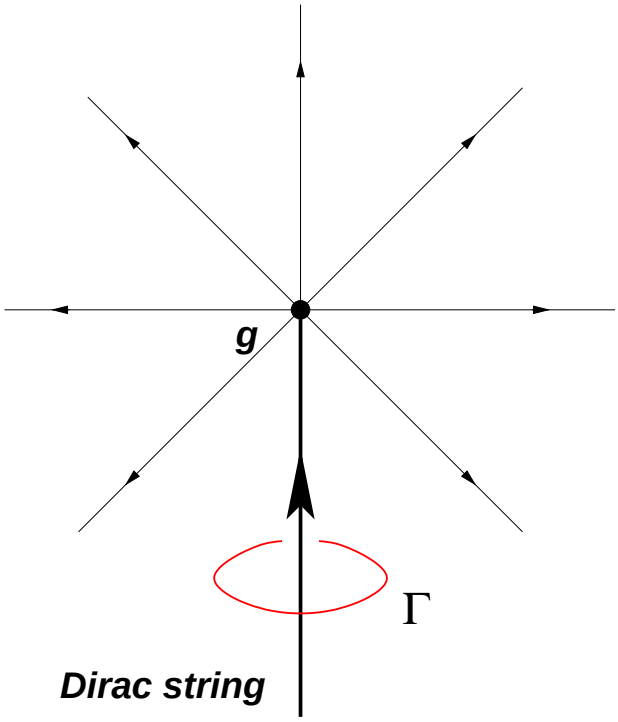


Fig. 8: The Dirac monopole.

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

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

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

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

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

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

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

$$e^{i\,e\,g} = 1 \qquad \Longrightarrow \qquad e\,g = 2\pi n \;\text{ with }\; n \in \mathbb{Z}. \tag{151}$$

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

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

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

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

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

$$\partial_\mu A^\mu = 0. \tag{153}$$

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

$$0 = \partial_\mu \partial^\mu A^\nu - \partial_\nu \left(\partial_\mu A^\mu \right) = \partial_\mu \partial^\mu A^\nu, \tag{154}$$

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

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

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

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

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

The canonical commutation relations imply that

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

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

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

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

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

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

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

This covariant derivative can be constructed in terms of the gauge potential A_μ as

$$D_\mu = \partial_\mu + ieA_\mu. \tag{159}$$

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

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

invariant under the gauge transformations

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

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

$$\mathcal{L}^{(\rm int)}_{\rm QED} = -eA_\mu\, \overline{\psi}\gamma^\mu\psi. \tag{162}$$

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

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

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

Let us consider a gauge group G with generators T^a, $a = 1, \ldots, \dim G$ satisfying the Lie algebra[6]

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

[6]Some basics facts about Lie groups have been summarized in Appendix A.

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

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

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

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

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

$$F_{\mu\nu} \longrightarrow U F_{\mu\nu} U^{-1}. \tag{166}$$

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

$$\Phi \longrightarrow U\Phi \tag{167}$$

the covariant derivative is given by

$$D_\mu \Phi = \partial_\mu \Phi - i g A_\mu^a T^a \Phi. \tag{168}$$

With the help of this we can write a generic Lagrangian for a nonabelian gauge field coupled to scalars ϕ and spinors ψ as

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

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

4.4 Understanding gauge symmetry

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

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

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

$$f_a(q, p) = 0, \qquad a = 1, \ldots, N_1. \tag{171}$$

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

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

$$\{f_a, H\} \approx 0. \tag{172}$$

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

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

$$\{\phi_a, \phi_b\} = c_{abc}\phi_c \approx 0. \tag{173}$$

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

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

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

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

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

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

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

Let us apply this to Maxwell electrodynamics described by the Lagrangian

$$L = -\frac{1}{4}\int d^3\, F_{\mu\nu}F^{\mu\nu}. \tag{177}$$

The generalized momentum conjugate to A_μ is given by

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

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

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

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

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

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

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

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

where we have absorbed A_0 in the definition of the arbitrary functions $\lambda_1(x)$ and $\lambda_2(x)$. Actually, we can fix part of the ambiguity taking $\lambda_1 = 0$. Notice that, because A_0 has been included in the multipliers, fixing λ_1 amounts to fixing the value of A_0 and therefore it is equivalent to taking a temporal gauge. In this case the Hamiltonian is

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

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

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

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

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

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

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

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

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

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

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

The role of gauge transformations in the quantum theory is very illuminating in understanding the real rôle of gauge invariance [25]. As we have learned, the existence of a gauge symmetry in a theory reflects a degree of redundancy in the description of physical states in terms of the degrees of freedom appearing in the Lagrangian. In Classical Mechanics, for example, the state of a system is usually determined by the value of the canonical coordinates (q_i, p_i). We know, however, that this is not the case for constrained Hamiltonian systems where the transformations generated by the first class constraints change the value of q_i and p_i withoug changing the physical state. In the case of Maxwell theory for every physical configuration determined by the gauge invariant quantities $\vec{E}$, $\vec{B}$ there is an infinite number of possible values of the vector potential that are related by gauge transformations $\delta A_\mu = \partial_\mu \varepsilon$.

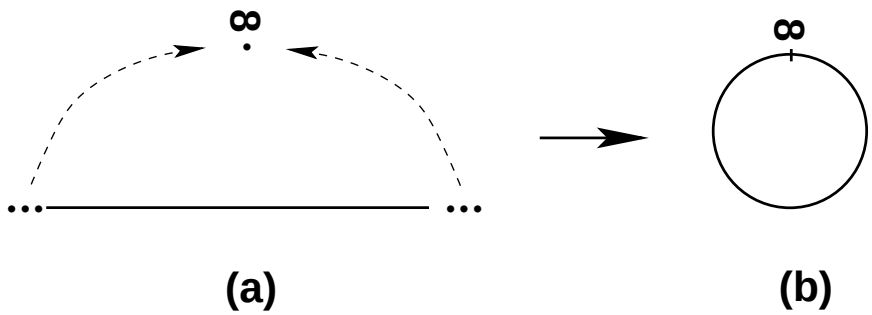


Fig. 9: Compactification of the real line (a) into the circumference S^1 (b) by adding the point at infinity.

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

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

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

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

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

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

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

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

$$\alpha(2\pi) = \alpha(0) + 2\pi\, n \ . \tag{189}$$

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

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

where the integral is along the spatial S^1.

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

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

[8] Although we present for simplicity only the case of SU(2), similar arguments apply to any simple group.

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

$$g : S^3 \longrightarrow S^3. \tag{192}$$

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

$$n = \frac{1}{24\pi^2}\int_{S^3} d^3x\, \epsilon_{ijk}\mathrm{Tr}\,\left[\left(g^{-1}\partial_i g\right)\left(g^{-1}\partial_i g\right)\left(g^{-1}\partial_i g\right)\right] \in \mathbb{Z}. \tag{193}$$

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

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

$$\mathcal{G}/\mathcal{G}_0 \neq \mathbf{1}. \tag{194}$$

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

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

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

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

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

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

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

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

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

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

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

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

However, it is possible to define a new vacuum state

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

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

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

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

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

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

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

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

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

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

Nevertheless a gauge invariant vacuum can be defined as

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

satisfying

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

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

$$\mathcal{L} \;=\; -\frac{1}{4}F^a_{\mu\nu}F^{\mu\nu\,a} - \frac{\theta g^2}{32\pi^2}F^a_{\mu\nu}\widetilde{F}^{\mu\nu\,a}, \tag{207}$$

where $\widetilde{F}^a_{\mu\nu}$ is the dual of the field strength defined by

$$\widetilde{F}^a_{\mu\nu} = \frac{1}{2}\epsilon_{\mu\nu\sigma\lambda}F^{\sigma\lambda}. \tag{208}$$

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

$$|\theta| < 10^{-9} \tag{209}$$

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

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

5 Towards computational rules: Feynman diagrams

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

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

5.1 Cross sections and S-matrix amplitudes

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

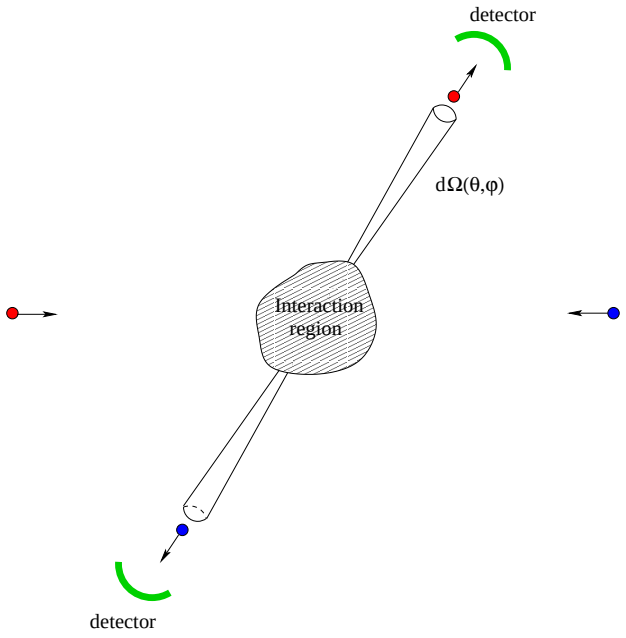


Fig. 10: Schematic setup of a two-to-two-particles single scattering event in the center of mass reference frame.

this quantity has to be proportional to the flux of incoming particles[9], $f_{\rm in}$. The proportionality constant defines the differential cross section

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

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

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

$$\sigma = \int_{-1}^{1} d(\cos\theta)\int_{0}^{2\pi} d\varphi\,\frac{d\sigma}{d\Omega}(\theta,\varphi). \tag{211}$$

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

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

$$|\vec{p}_1,\ldots,\vec{p}_n;{\rm in}\rangle. \tag{212}$$

[9]This is defined as the number of particles that enter the interaction region per unit time and per unit area perpendicular to the direction of the beam.

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

$$|\vec{p}_1{}', \ldots, \vec{p}_k{}'; \text{out}\rangle \tag{213}$$

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

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

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

$$\langle \vec{p}_1{}', \ldots, \vec{p}_k{}'; \text{out} | \vec{p}_1, \ldots, \vec{p}_n; \text{in}\rangle = \langle \vec{p}_1{}', \ldots, \vec{p}_k{}' | \widehat{S} | \vec{p}_1, \ldots, \vec{p}_n\rangle. \tag{215}$$

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

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

$$\widehat{S} = \mathbf{1} + i\widehat{T}, \tag{216}$$

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

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

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

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

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

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

5.2 Feynman rules

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

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

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

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

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

$$\begin{gathered}\langle\vec{p}_1{}',\dots,\vec{p}_{m'}{}';\mathrm{out}|\vec{p}_1,\dots,\vec{p}_n;\mathrm{in}\rangle \\ \Downarrow \\ \int d^4x_1\dots\int d^4y_n\langle\Omega|T\Big[\phi(x_1)^\dagger\dots\phi(x_m)^\dagger\phi(y_1)\dots\phi(y_n)\Big]|\Omega\rangle\, e^{ip_1{}'\cdot x_1}\dots e^{-ip_n\cdot y_n},\end{gathered} \tag{221}$$

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

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

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

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

$$S[\phi,\phi^\dagger] = S_0[\phi,\phi^\dagger] + S_{\mathrm{int}}[\phi,\phi^\dagger]. \tag{223}$$

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

$$\int\mathscr{D}\phi\mathscr{D}\phi^\dagger\Big[\dots\Big]e^{iS_0[\phi,\phi^\dagger]}, \tag{224}$$

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

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

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

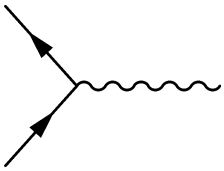
For the sake of concreteness we focus on the case of QED first. Going back to Eq. (160) we expand the covariant derivative to write the action

$$S_{\rm QED} = \int d^4x \left[-\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \overline{\psi}(i\not{\partial} - m)\psi + e\overline{\psi}\gamma^\mu \psi A_\mu \right] . \tag{225}$$

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

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

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



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

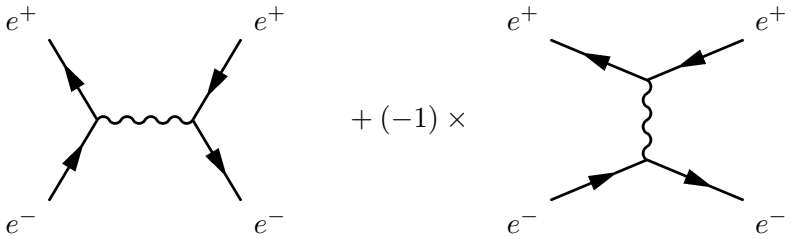
To show in a practical way how Feynman diagrams are drawn, we consider Bhabha scattering, i.e. the elastic dispersion of an electron and a positron:

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

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

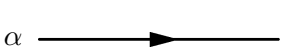
[10]From the point of view of the operator formalism, the requirement of considering only diagrams that are topologically nonequivalent comes from the fact that each diagram represents a certain Wick contraction in the correlation function of interaction-picture operators.

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



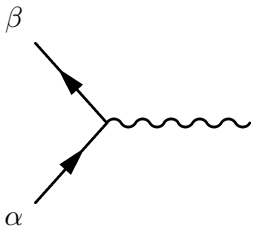
Incoming and outgoing particles appear respectively on the left and the right of this diagram. Notice how the identification of electrons and positrons is done comparing the direction of the charge flux with the direction of propagation. For electrons the flux of charges goes in the direction of propagation,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

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



$$\Longrightarrow \qquad \left(\frac{i}{\not{p} - m + i\varepsilon}\right)_{\beta\alpha}$$

$$\Longrightarrow \qquad \frac{-i\eta_{\mu\nu}}{p^2 + i\varepsilon}$$



$$\Longrightarrow \qquad -ie\gamma^{\mu}_{\beta\alpha}(2\pi)^4\delta^{(4)}(p_1 + p_2 + p_3).$$

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

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

and introduce a factor of -1 for each fermion loop in the diagram[11].

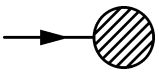
[11] The contribution of each diagram comes also multiplied by a degeneracy factor that takes into account in how many ways a given Wick contraction can be done. In QED, however, these factors are equal to 1 for many diagrams.

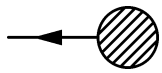
In fact, some of the integrations over internal momenta can actually be done using the delta function at the vertices, leaving just a global delta function implementing the total momentum conservation in the diagram [cf. Eq. (217)]. It is even possible that all integrations can be eliminated in this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

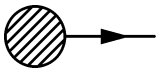
The need to perform integrations over internal momenta in loop diagrams has important consequences in Quantum Field Theory. The reason is that in many cases the resulting integrals are ill-defined, i.e. are divergent either at small or large values of the loop momenta. In the first case one speaks of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

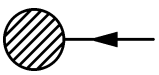
Were we computing time-ordered (amputated) correlation function of operators, this would be all. However, in the case of S-matrix amplitudes this is not the whole story. In addition to the previous rules here one needs to attach contributions also to the external legs in the diagram. These are the wave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

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

Incoming fermion: α  $\Longrightarrow$ $u_\alpha(\vec{p}, s)$

Incoming antifermion: α  $\Longrightarrow$ $\overline{v}_\alpha(\vec{p}, s)$

Outgoing fermion:  α $\Longrightarrow$ $\overline{u}_\alpha(\vec{p}, s)$

Outgoing antifermion:  α $\Longrightarrow$ $v_\alpha(p, s)$

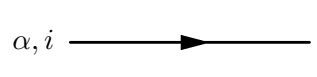
Incoming photon: μ  $\Longrightarrow$ $\epsilon_\mu(\vec{k}, \lambda)$

Outgoing photon:  $\Longrightarrow \quad \epsilon_\mu(\vec{k},\lambda)^*$

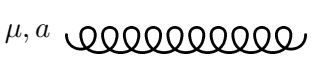
Here we have assumed that the momenta for incoming (resp. outgoing) particles are entering (resp. leaving) the diagram. It is important also to keep in mind that in the computation of S-matrix amplitudes all external states are on-shell. In Section 5.3 we illustrate the use of the Feynman rules for QED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

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

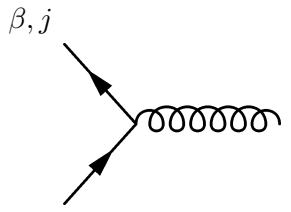
Above we have illustrated the Feynman rules with the case of QED. Similar rules can be computed for other interacting quantum field theories with scalar, vector or spinor fields. In the case of the nonabelian gauge theories introduced in Section 4.3 we have:



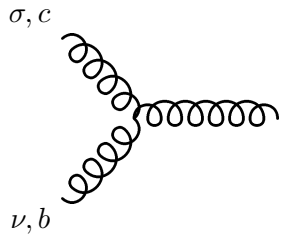
$\Longrightarrow \quad \left(\dfrac{i}{\not{p} - m + i\varepsilon}\right)_{\beta\alpha} \delta_{ij}$



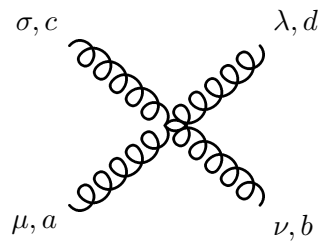
$\Longrightarrow \quad \dfrac{-i\eta_{\mu\nu}}{p^2 + i\varepsilon}\delta^{ab}$



$\Longrightarrow \quad -ig\gamma^\mu_{\beta\alpha} t^a_{ij}$



$\Longrightarrow \quad g\, f^{abc}\Big[\eta^{\mu\nu}(p_1^\sigma - p_2^\sigma) + \text{permutations}\Big]$



$\Longrightarrow \quad -ig^2\Big[f^{abe}f^{cde}\Big(\eta^{\mu\sigma}\eta^{\nu\lambda} - \eta^{\mu\lambda}\eta^{\nu\sigma}\Big) + \text{permutations}\Big]$

It is not our aim here to give a full and detailed description of the Feynman rules for nonabelian gauge theories. It suffices to point out that, unlike the case of QED, here the gauge fields can interact among themselves. Indeed, the three and four gauge field vertices are a consequence of the cubic and quartic terms in the action

$$S = -\frac{1}{4}\int d^4x\, F^a_{\mu\nu}F^{\mu\nu\,a}, \tag{227}$$

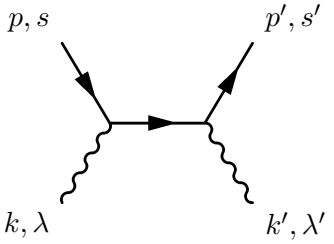
where the nonabelian gauge field strength $F^a_{\mu\nu}$ is given in Eq. (165). The self-interaction of the nonabelian gauge fields has crucial dynamical consequences and its at the very heart of its success in describing the physics of elementary particles.

5.3 An example: Compton scattering

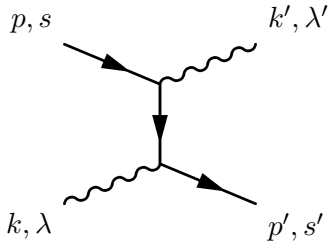
To illustrate the use of Feynman diagrams and Feynman rules we compute the cross section for the dispersion of photons by free electrons, the so-called Compton scattering:

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

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

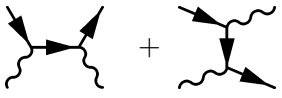


This is, however, not the only possibility. Indeed, there is a second possible diagram:



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

Using the Feynman rules of QED we find



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

Because the leading order contributions only involve tree-level diagrams, there is no integration over internal momenta and therefore we are left with a purely algebraic expression for the amplitude. To get an explicit expression we begin by simplifying the numerators. The following simple identity turns out to be very useful for this task

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

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

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

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

$$|\vec{p}|, |\vec{k}|, |\vec{p}\,'|, |\vec{k}\,'| \ll m_e. \tag{231}$$

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

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

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

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

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

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

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

$$p^2 = m_e^2 = p'^2 \qquad \text{and} \qquad k^2 = 0 = k'^2. \tag{235}$$

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

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

and

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

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

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

Putting together all these expressions we find that at low energies

$$\approx \frac{(ie)^2}{2m_e}\overline{u}(\vec{p}\,',s')\left[\not{\epsilon}'(\vec{k}\,'\lambda')^*\frac{\not{k}}{|\vec{k}|}\epsilon(\vec{k},\lambda) + \epsilon(\vec{k},\lambda)\frac{\not{k}'}{|\vec{k}\,'|}\not{\epsilon}'(\vec{k}\,'\lambda')^*\right]u(\vec{p},s). \tag{239}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\sum_{s=\pm\frac{1}{2}}\sum_{s'=\pm\frac{1}{2}}\Big|\overline{u}(\vec{p}\,',s')\not{k}u(\vec{p},s)\Big|^2=\sum_{s=\pm\frac{1}{2}}\sum_{s'=\pm\frac{1}{2}}\Big[u(\vec{p},s)^\dagger\not{k}^\dagger\overline{u}(\vec{p}\,',s')^\dagger\Big]\Big[\overline{u}(\vec{p}\,',s')\not{k}u(\vec{p},s)\Big], \tag{245}$$

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

$$\sum_{s=\pm\frac{1}{2}}\sum_{s'=\pm\frac{1}{2}}\left|\overline{u}(\vec{p}',s')\not{k}u(\vec{p},s)\right|^2 = \left[\sum_{s=\pm\frac{1}{2}}u_\alpha(\vec{p},s)\overline{u}_\beta(\vec{p},s)\right](\not{k})_{\beta\sigma}\left[\sum_{s'=\pm\frac{1}{2}}u_\sigma(\vec{p}',s')\overline{u}_\rho(\vec{p}',s')\right](\not{k})_{\rho\alpha}$$
$$= \mathrm{Tr}\left[(\not{p}+m_e)\not{k}(\not{p}'+m_e)\not{k}\right], \tag{246}$$

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

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

This gives the following value for the invariant amplitude

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

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

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

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

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

where σ_T is the total Thomson cross section

$$\sigma_T = \frac{e^4}{6\pi m_e^2} = \frac{8\pi}{3}r_{\mathrm{cl}}^2. \tag{251}$$

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

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

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

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

[12] We use also the fact that the trace of the product of an odd number of Dirac matrices is always zero.

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

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

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

6 Symmetries

6.1 Noether's theorem

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

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

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

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

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

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

$$\delta_\epsilon \mathcal{L} = \partial_\mu K^\mu. \tag{256}$$

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

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

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

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

[13]The following result can be also derived a more general situations where the Lagrangian changes by a total time derivative.

Actually a conserved current implies the existence of a charge

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

which is conserved

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

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

$$\delta\phi = i[\phi, Q]. \tag{261}$$

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

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

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

$$\delta\mathcal{L} = -a^\mu \partial_\mu \mathcal{L}. \tag{263}$$

Therefore the corresponding conserved charge is

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

where we introduced the energy-momentum tensor

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

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

$$\partial_\mu T^{\mu\nu} = 0. \tag{266}$$

The associated conserved charges are given by

$$P^\nu = \int d^3x\, T^{0\nu} \tag{267}$$

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

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

$$j^\mu = \overline{\psi}\gamma^\mu\psi, \qquad \partial_\mu j^\mu = 0. \tag{268}$$

Thus implying the existence of a conserved charge

$$Q = \int d^3x \overline{\psi}\gamma^0\psi = \int d^3x \psi^\dagger\psi. \tag{269}$$

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

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

$$\psi_i \longrightarrow U_{ij}\psi_j, \qquad U^\dagger U = \mathbf{1} \tag{270}$$

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

$$U = \exp\left(i\alpha_a T^a\right), \qquad (T^a)^\dagger = T^a, \tag{271}$$

we find the conserved current

$$j^{\mu\,a} = \overline{\psi}_i T^a_{ij}\gamma^\mu\psi_j, \qquad \partial_\mu j^\mu = 0. \tag{272}$$

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

$$\begin{pmatrix} u \\ d \end{pmatrix} \longrightarrow M \begin{pmatrix} u \\ d \end{pmatrix}, \qquad M \in \mathrm{SU}(2). \tag{273}$$

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

6.2 Symmetries in the quantum theory

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

$$[Q^a, H] = 0. \tag{274}$$

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

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

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

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

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

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

$$|i\rangle = \phi_i|0\rangle \tag{277}$$

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

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

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

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

$$e^{i\alpha_a Q^a}|0\rangle \neq |0\rangle \qquad \Longrightarrow \qquad Q^a|0\rangle \neq 0. \tag{279}$$

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

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

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

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

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

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

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

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

and look for the minimum of

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

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

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

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

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

[14]For simplicity we consider that the minima of $V(\phi)$ occur at zero potential.

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

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

$$(K^A)^i_j\langle\varphi^j\rangle \neq 0. \tag{286}$$

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

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

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

and the mass matrix is:

$$M^2_{ij} \equiv \left.\frac{\partial^2 V}{\partial\varphi^i\partial\varphi^j}\right|_{\varphi=\langle\varphi\rangle}. \tag{288}$$

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

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

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

Differentiating this expression with respect to φ^k we arrive at

$$\frac{\partial^2 V}{\partial\varphi^i\partial\varphi^k}(T^a)^i_j\varphi^j + \frac{\partial V}{\partial\varphi^i}(T^a)^i_k = 0. \tag{290}$$

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

$$M^2_{ik}(T^a)^i_j\langle\varphi^j\rangle = 0. \tag{291}$$

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

$$M^2_{ik}(K^A)^i_j\langle\varphi^j\rangle = 0. \tag{292}$$

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

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

$$V(\vec{\varphi}) = \frac{\lambda}{4}\left(\vec{\varphi}^{\,2} - a^2\right)^2 . \tag{293}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\Phi \to e^{i\alpha(x)} \Phi, \qquad A_\mu \to A_\mu + \partial_\mu \alpha(x). \tag{297}$$

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

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

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

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

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

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

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

$$m_\gamma^2 = 2 e^2 \mu^2 . \tag{301}$$

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

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

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

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

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

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

The Abelian Higgs model discussed here can be regarded as a toy model of the Higgs mechanism responsible for giving mass to the $W^\pm$ and Z^0 gauge bosons in the Standard Model. In condensed matter physics the symmetry breaking described by the nonrelativistic version of the Abelian Higgs model can be used to characterize the onset of a superconducting phase in the BCS theory, where the complex scalar field Φ is associated with the Cooper pairs. In this case the parameter μ^2 depends on the temperature. Above the critical temperature T_c, $\mu^2(T)>0$ and there is only a symmetric vacuum $\langle\Phi\rangle=0$. When, on the other hand, $T<T_c$ then $\mu^2(T)<0$ and symmetry breaking takes place. The onset of a nonzero

mass of the photon (301) below the critical temperature explains the Meissner effect: the magnetic fields cannot penetrate inside superconductors beyond a distance of the order $\frac{1}{m_\gamma}$.

7 Anomalies

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

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

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

7.1 Axial anomaly

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

$$\mathcal{L} = i\overline{\psi}\partial\!\!\!/\psi = iu_+^\dagger \sigma_+^\mu \partial_\mu u_+ + iu_-^\dagger \sigma_-^\mu \partial_\mu u_- \quad \text{with} \qquad \psi = \begin{pmatrix} u_+ \\ u_- \end{pmatrix} \tag{303}$$

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

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

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

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

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

$$J_V^\mu = \overline{\psi}\gamma^\mu\psi = u_+^\dagger \sigma_+^\mu u_+ + u_-^\dagger \sigma_-^\mu u_- \quad \Longrightarrow \quad \partial_\mu J_V^\mu = 0 \tag{306}$$

and an axial vector current

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

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

$$\partial_\mu J_A^\mu \sim \hbar\, \vec{E}\cdot\vec{B}. \tag{308}$$

To understand more clearly how this result comes about we study first a simple model in two dimensions that captures the relevant physics involved in the four-dimensional case [29]. We work in

Minkowski space in two dimensions with coordinates $(x^0, x^1) \equiv (t, x)$ and where the spatial direction is compactified to a circle S^1. In this setup we consider a fermion coupled to the electromagnetic field. Notice that since we are living in two dimensions the field strength $F_{\mu\nu}$ only has one independent component that corresponds to the electric field along the spatial direction, $F^{01} \equiv \mathcal{E}$ (in two dimensions there are no magnetic fields!).

To write the Lagrangian for the spinor field we need to find a representation of the algebra of γ-matrices

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu} \quad \text{with} \quad \eta = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{309}$$

In two dimensions the dimension of the representation of the γ-matrices is $2^{[\frac{2}{2}]} = 2$. Here take

$$\gamma^0 \equiv \sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \gamma^1 \equiv i\sigma^2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{310}$$

This is a chiral representation since the matrix γ_5 is diagonal[15]

$$\gamma_5 \equiv -\gamma^0\gamma^1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{311}$$

Writing the two-component spinor ψ as

$$\psi = \begin{pmatrix} u_+ \\ u_- \end{pmatrix} \tag{312}$$

and defining as usual the projectors $P_\pm = \frac{1}{2}(\mathbf{1} \pm \gamma_5)$ we find that the components $u_\pm$ of ψ are respectively a right- and left-handed Weyl spinor in two dimensions.

Once we have a representation of the γ-matrices we can write the Dirac equation. Expressing it in terms of the components $u_\pm$ of the Dirac spinor we find

$$(\partial_0 - \partial_1)u_+ = 0, \qquad (\partial_0 + \partial_1)u_- = 0. \tag{313}$$

The general solution to these equations can be immediately written as

$$u_+ = u_+(x^0 + x^1), \qquad u_- = u_-(x^0 - x^1). \tag{314}$$

Hence $u_\pm$ are two wave packets moving along the spatial dimension respectively to the left (u_+) and to the right (u_-). Notice that according to our convention the left-moving u_+ is a right-handed spinor (positive helicity) whereas the right-moving u_- is a left-handed spinor (negative helicity).

If we want to interpret (313) as the wave equation for two-dimensional Weyl spinors we have the following wave functions for free particles with well defined momentum $p^\mu = (E, p)$.

$$u_\pm^{(E)}(x^0 \pm x^1) = \frac{1}{\sqrt{L}} e^{-iE(x^0 \pm x^1)} \quad \text{with} \quad p = \mp E. \tag{315}$$

As it is always the case with the Dirac equation we have both positive and negative energy solutions. For u_+, since $E = -p$, we see that the solutions with positive energy are those with negative momentum $p < 0$, whereas the negative energy solutions are plane waves with $p > 0$. For the left-handed spinor u_- the situation is reversed. Besides, since the spatial direction is compact with length L the momentum p is quantized according to

$$p = \frac{2\pi n}{L}, \qquad n \in \mathbb{Z}. \tag{316}$$

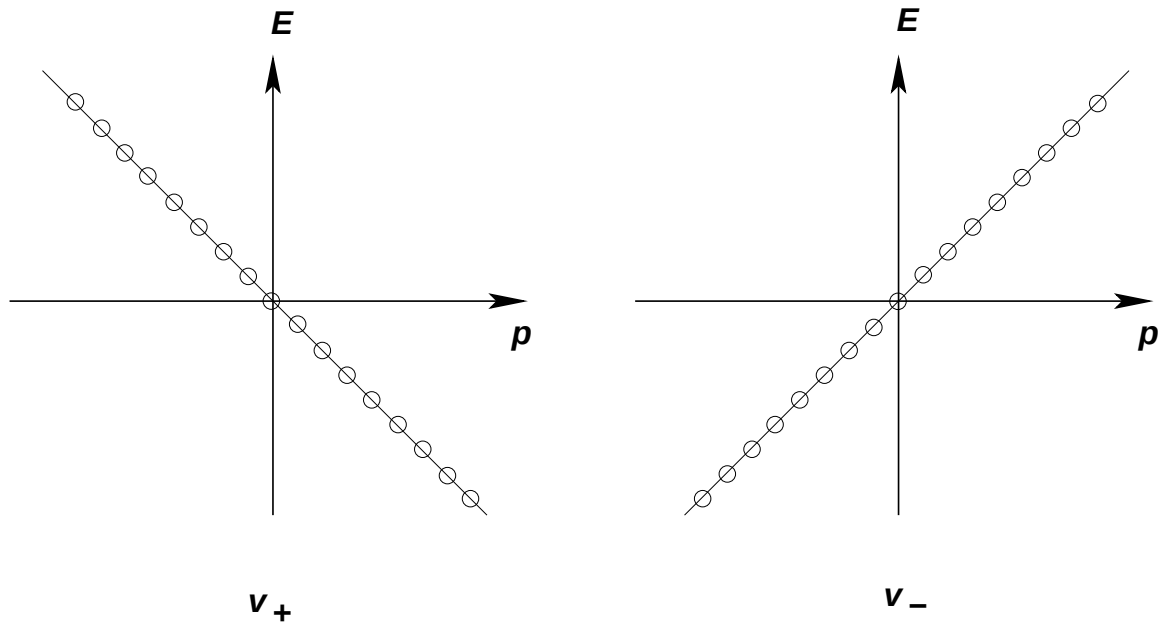


Fig. 11: Spectrum of the massless two-dimensional Dirac field.

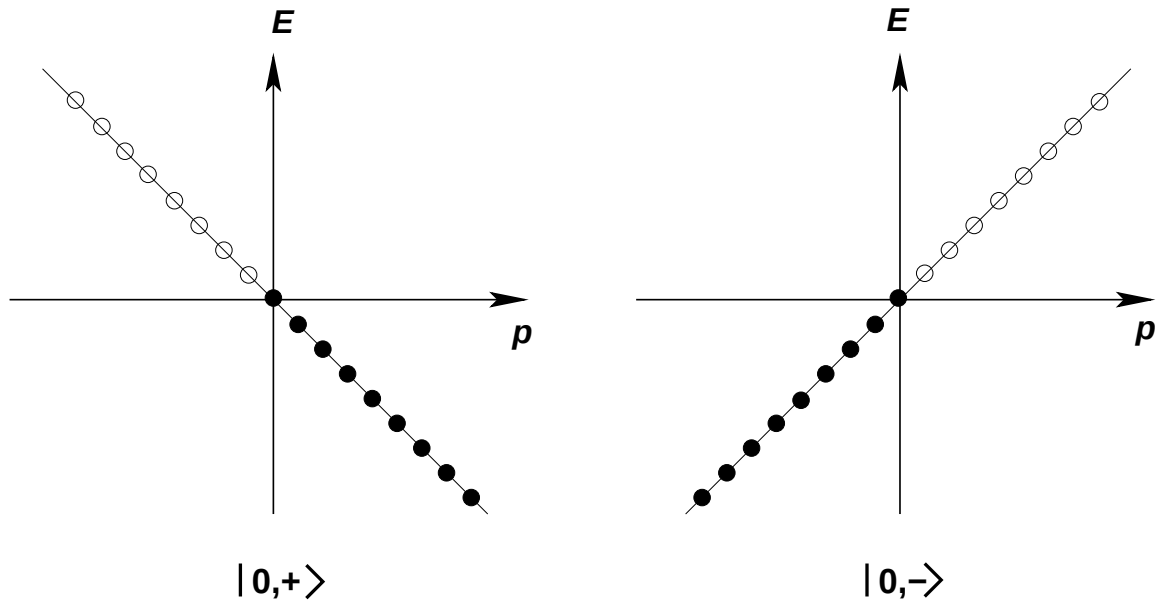


Fig. 12: Vacuum of the theory.

The spectrum of the theory is represented in Fig. 11.

Once we have the spectrum of the theory the next step is to obtain the vacuum. As with the Dirac equation in four dimensions we fill all the states with $E \leq 0$ (Fig. 12). Exciting of a particle in the Dirac see produces a positive energy fermion plus a hole that is interpreted as an antiparticle. This gives us the clue on how to quantize the theory. In the expansion of the operator $u_{\pm}$ in terms of the modes (315) we associate positive energy states with annihilation operators whereas the states with negative energy are

[15] In any even number of dimensions γ_5 is defined to satisfy the conditions $\gamma_5^2 = \mathbf{1}$ and $\{\gamma_5, \gamma^\mu\} = 0$.

associated with creation operators for the corresponding antiparticle

$$u_\pm(x) = \sum_{E>0} \left[a_\pm(E) v_\pm^{(E)}(x) + b_\pm^\dagger(E) v_\pm^{(E)}(x)^* \right]. \tag{317}$$

The operator $a_\pm(E)$ acting on the vacuum $|0,\pm\rangle$ annihilates a particle with positive energy E and momentum $\mp E$. In the same way $b_\pm^\dagger(E)$ creates out of the vacuum an antiparticle with positive energy E and spatial momentum $\mp E$. In the Dirac sea picture the operator $b_\pm(E)^\dagger$ is originally an annihilation operator for a state of the sea with negative energy $-E$. As in the four-dimensional case the problem of the negative energy states is solved by interpreting annihilation operators for negative energy states as creation operators for the corresponding antiparticle with positive energy (and vice versa). The operators appearing in the expansion of $u_\pm$ in Eq. (317) satisfy the usual algebra

$$\{a_\lambda(E), a_{\lambda'}^\dagger(E')\} = \{b_\lambda(E), b_{\lambda'}^\dagger(E')\} = \delta_{E,E'}\delta_{\lambda\lambda'}, \tag{318}$$

where we have introduced the label $\lambda, \lambda' = \pm$. Also, $a_\lambda(E)$, $a_\lambda^\dagger(E)$ anticommute with $b_{\lambda'}(E')$, $b_{\lambda'}^\dagger(E')$.

The Lagrangian of the theory

$$\mathcal{L} = i u_+^\dagger(\partial_0 + \partial_1) u_+ + i u_-^\dagger(\partial_0 - \partial_1) u_- \tag{319}$$

is invariant under both U(1)$_V$, Eq. (304), and U(1)$_A$, Eq. (305). The associated Noether currents are in this case

$$J_V^\mu = \begin{pmatrix} u_+^\dagger u_+ + u_-^\dagger u_- \\ -u_+^\dagger u_+ + u_-^\dagger u_- \end{pmatrix}, \qquad J_A^\mu = \begin{pmatrix} u_+^\dagger u_+ - u_-^\dagger u_- \\ -u_+^\dagger u_+ - u_-^\dagger u_- \end{pmatrix}. \tag{320}$$

The associated conserved charges are given, for the vector current by

$$Q_V = \int_0^L dx^1 \left(u_+^\dagger u_+ + u_-^\dagger u_- \right) \tag{321}$$

and for the axial current

$$Q_A = \int_0^L dx^1 \left(u_+^\dagger u_+ - u_-^\dagger u_- \right). \tag{322}$$

Using the orthonormality relations for the modes $v_\pm^{(E)}(x)$

$$\int_0^L dx^1 \, v_\pm^{(E)}(x) \, v_\pm^{(E')}(x) = \delta_{E,E'} \tag{323}$$

we find for the conserved charges:

$$\begin{aligned} Q_V &= \sum_{E>0} \left[a_+^\dagger(E) a_+(E) - b_+^\dagger(E) b_+(E) + a_-^\dagger(E) a_-(E) - b_-^\dagger(E) b_-(E) \right], \\ Q_A &= \sum_{E>0} \left[a_+^\dagger(E) a_+(E) - b_+^\dagger(E) b_+(E) - a_-^\dagger(E) a_-(E) + b_-^\dagger(E) b_-(E) \right]. \end{aligned} \tag{324}$$

We see that Q_V counts the net number (particles minus antiparticles) of positive helicity states plus the net number of states with negative helicity. The axial charge, on the other hand, counts the net number of positive helicity states minus the number of negative helicity ones. In the case of the vector current we have subtracted a formally divergent vacuum contribution to the charge (the "charge of the Dirac sea").

In the free theory there is of course no problem with the conservation of either Q_V or Q_A, since the occupation numbers do not change. What we want to study is the effect of coupling the theory to electric

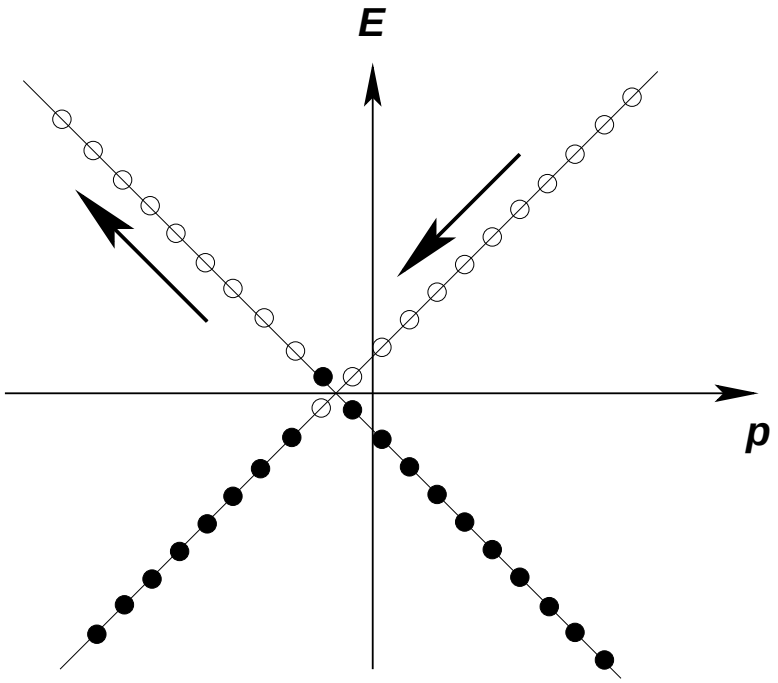


Fig. 13: Effect of the electric field.

field $\mathcal{E}$. We work in the gauge $A_0 = 0$. Instead of solving the problem exactly we are going to simulate the electric field by adiabatically varying in a long time τ_0 the vector potential A_1 from zero value to $-\mathcal{E}\tau_0$. From our discussion in section 4.3 we know that the effect of the electromagnetic coupling in the theory is a shift in the momentum according to

$$p \longrightarrow p - eA_1, \tag{325}$$

where e is the charge of the fermions. Since we assumed that the vector potential varies adiabatically, we can assume it to be approximately constant at each time.

Then, we have to understand what is the effect of (325) on the vacuum depicted in Fig. (12). What we find is that the two branches move as shown in Fig. (13) resulting in some of the negative energy states of the v_+ branch acquiring positive energy while the same number of the empty positive energy states of the other branch v_- will become empty negative energy states. Physically this means that the external electric field $\mathcal{E}$ creates a number of particle-antiparticle pairs out of the vacuum. Denoting by $N \sim e\mathcal{E}$ the number of such pairs created by the electric field per unit time, the final values of the charges Q_V and Q_A are

$$\begin{aligned} Q_A(\tau_0) &= (N-0) + (0-N) = 0, \\ Q_V(\tau_0) &= (N-0) - (0-N) = 2N. \end{aligned} \tag{326}$$

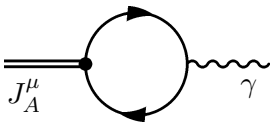
Therefore we conclude that the coupling to the electric field produces a violation in the conservation of the axial charge per unit time given by $\Delta Q_A \sim e\mathcal{E}$. This implies that

$$\partial_\mu J^\mu_A \sim e\hbar\mathcal{E}, \tag{327}$$

where we have restored $\hbar$ to make clear that the violation in the conservation of the axial current is a quantum effect. At the same time $\Delta Q_V = 0$ guarantees that the vector current remains conserved also quantum mechanically, $\partial_\mu J^\mu_V = 0$.

We have just studied a two-dimensional example of the Adler-Bell-Jackiw axial anomaly [30]. The heuristic analysis presented here can be made more precise by computing the quantity

$$C^{\mu\nu} = \langle 0|T\left[J_A^\mu(x)J_V^\nu(0)\right]|0\rangle = \tag{328}$$



The anomaly is given then by $\partial_\mu C^{\mu\nu}$. A careful calculation yields the numerical prefactor missing in Eq. (327) leading to the result

$$\partial_\mu J_A^\mu = \frac{e\hbar}{2\pi}\varepsilon^{\nu\sigma}F_{\nu\sigma}, \tag{329}$$

with $\varepsilon^{01} = -\varepsilon^{10} = 1$.

The existence of an anomaly in the axial symmetry that we have illustrated in two dimensions is present in all even dimensional of space-times. In particular in four dimensions the axial anomaly it is given by

$$\partial_\mu J_A^\mu = -\frac{e^2}{16\pi^2}\varepsilon^{\mu\nu\sigma\lambda}F_{\mu\nu}F_{\sigma\lambda}. \tag{330}$$

This result has very important consequences in the physics of strong interactions as we will see in what follows

7.2 Chiral symmetry in QCD

Our knowledge of the physics of strong interactions is based on the theory of Quantum Chromodynamics (QCD) [32]. This is a nonabelian gauge theory with gauge group SU(N_c) coupled to a number N_f of quarks. These are spin-$\frac{1}{2}$ particles $Q^{i\,f}$ labelled by two quantum numbers: color $i = 1, \ldots, N_c$ and flavor $f = 1, \ldots, N_f$. The interaction between them is mediated by the $N_c^2 - 1$ gauge bosons, the gluons A_μ^a, $a = 1, \ldots, N_c^2 - 1$. In the real world $N_c = 3$ and the number of flavors is six, corresponding to the number of different quarks: up (u), down (d), charm (c), strange (s), top (t) and bottom (b).

For the time being we are going to study a general theory of QCD with N_c colors and N_f flavors. Also, for reasons that will be clear later we are going to work in the limit of vanishing quark masses, $m_f \to 0$. In this cases the Lagrangian is given by

$$\mathcal{L}_{\rm QCD} = -\frac{1}{4}F^a_{\mu\nu}F^{a\,\mu\nu} + \sum_{f=1}^{N_f}\left[i\overline{Q}^f_L \not{D}\, Q^f_L + i\overline{Q}^f_R \not{D}\, Q^f_R\right], \tag{331}$$

where the subscripts L and R indicate respectively left and right-handed spinors, $Q^f_{L,R} \equiv P_\pm Q^f$, and the field strength $F^a_{\mu\nu}$ and the covariant derivative D_μ are respectively defined in Eqs. (165) and (168). Apart from the gauge symmetry, this Lagrangian is also invariant under a global U(N_f)$_L$×U(N_f)$_R$ acting on the flavor indices and defined by

$$\mathrm{U}(N_f)_L : \begin{cases} Q^f_L \;\to\; \sum_{f'}(U_L)_{ff'}Q^{f'}_L \\ Q^f_R \;\to\; Q^f_R \end{cases} \qquad \mathrm{U}(N_f)_R : \begin{cases} Q^f_L \;\to\; Q^f_L \\ Q^r_R \;\to\; \sum_{f'}(U_R)_{ff'}Q^{f'}_R \end{cases} \tag{332}$$

with $U_L, U_R \in \mathrm{U}(N_f)$. Actually, since U($N$)=U(1)×SU($N$) this global symmetry group can be written as $\mathrm{SU}(N_f)_L \times \mathrm{SU}(N_f)_R \times \mathrm{U}(1)_L \times \mathrm{U}(1)_R$. The abelian subgroup $\mathrm{U}(1)_L \times \mathrm{U}(1)_R$ can be now decomposed into their vector $\mathrm{U}(1)_B$ and axial $\mathrm{U}(1)_A$ subgroups defined by the transformations

$$\mathrm{U}(1)_B : \begin{cases} Q_L^f & \to & e^{i\alpha} Q_L^f \\ \\ Q_R^f & \to & e^{i\alpha} Q_R^f \end{cases} \qquad \mathrm{U}(1)_A : \begin{cases} Q_L^f & \to & e^{i\alpha} Q_L^f \\ \\ Q_R^f & \to & e^{-i\alpha} Q_R^f \end{cases} \tag{333}$$

According to Noether's theorem, associated with these two abelian symmetries we have two conserved currents:

$$J_V^\mu = \sum_{f=1}^{N_f} \overline{Q}^f \gamma^\mu \, Q^f, \qquad J_A^\mu = \sum_{f=1}^{N_f} \overline{Q}^f \gamma^\mu \gamma_5 \, Q^f. \tag{334}$$

The conserved charge associated with vector charge J_V^μ is actually the baryon number defined as the number of quarks minus number of antiquarks.

The nonabelian part of the global symmetry group $\mathrm{SU}(N_f)_L \times \mathrm{SU}(N_f)_R$ can also be decomposed into its vector and axial subgroups, $\mathrm{SU}(N_f)_V \times \mathrm{SU}(N_f)_A$, defined by the following transformations of the quarks fields

$$\mathrm{SU}(N_f)_V : \begin{cases} Q_L^f & \to & \sum_{f'} (U_L)_{ff'} Q_L^{f'} \\ \\ Q_R^f & \to & \sum_{f'} (U_L)_{ff'} Q_R^{f'} \end{cases} \qquad \mathrm{SU}(N_f)_A : \begin{cases} Q_L^f & \to & \sum_{f'} (U_L)_{ff'} Q_L^{f'} \\ \\ Q_R^f & \to & \sum_{f'} (U_R^{-1})_{ff'} Q_R^{f'} \end{cases} \tag{335}$$

Again, the application of Noether's theorem shows the existence of the following nonabelian conserved charges

$$J_V^{I\,\mu} \equiv \sum_{f,f'=1}^{N_f} \overline{Q}^f \gamma^\mu (T^I)_{ff'} Q^{f'}, \qquad J_A^{I\,\mu} \equiv \sum_{f,f'=1}^{N_f} \overline{Q}^f \gamma^\mu \gamma_5 (T^I)_{ff'} Q^{f'}. \tag{336}$$

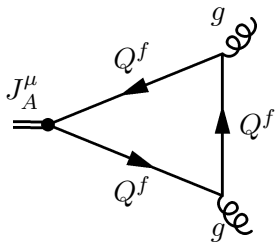
To summarize, we have shown that the initial chiral symmetry of the QCD Lagrangian (331) can be decomposed into its chiral and vector subgroups according to

$$\mathrm{U}(N_f)_L \times \mathrm{U}(N_f)_R = \mathrm{SU}(N_f)_V \times \mathrm{SU}(N_f)_A \times \mathrm{U}(1)_B \times \mathrm{U}(1)_A. \tag{337}$$

The question to address now is which part of the classical global symmetry is preserved by the quantum theory.

As argued in section 7.1, the conservation of the axial currents J_A^μ and $J_A^{a\,\mu}$ can in principle be spoiled due to the presence of an anomaly. In the case of the abelian axial current J_A^μ the relevant quantity is the correlation function

$$C^{\mu\nu\sigma} \equiv \langle 0 | T \left[J_A^\mu(x) j_{\mathrm{gauge}}^{a\,\nu}(x') j_{\mathrm{gauge}}^{b\,\sigma}(0) \right] |0\rangle = \sum_{f=1}^{N_f} \Bigg[\quad \Bigg]_{\mathrm{symmetric}} \tag{338}$$



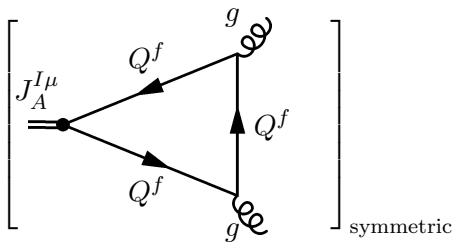
Here $j_{\mathrm{gauge}}^{a\,\mu}$ is the nonabelian conserved current coupling to the gluon field

$$j_{\mathrm{gauge}}^{a\,\mu} \equiv \sum_{f=1}^{N_f} \overline{Q}^f \gamma^\mu \tau^a Q^f, \tag{339}$$

where, to avoid confusion with the generators of the global symmetry we have denoted by τ^a the generators of the gauge group SU(N_c). The anomaly can be read now from $\partial_\mu C^{\mu\nu\sigma}$. If we impose Bose symmetry with respect to the interchange of the two outgoing gluons and gauge invariance of the whole expression, $\partial_\nu C^{\mu\nu\sigma} = 0 = \partial_\sigma C^{\mu\nu\sigma}$, we find that the axial abelian global current has an anomaly given by[16]

$$\partial_\mu J_A^\mu = -\frac{g^2 N_f}{32\pi^2}\varepsilon^{\mu\nu\sigma\lambda} F^a_{\mu\nu} F^{a\,\mu\nu}. \tag{340}$$

In the case of the nonabelian axial global symmetry SU(N_f)$_A$ the calculation of the anomaly is made as above. The result, however, is quite different since in this case we conclude that the nonabelian axial current $J_A^{a\,\mu}$ is not anomalous. This can be easily seen by noticing that associated with the axial current vertex we have a generator T^I of SU(N_f), whereas for the two gluon vertices we have the generators τ^a of the gauge group SU(N_c). Therefore, the triangle diagram is proportional to the group-theoretic factor

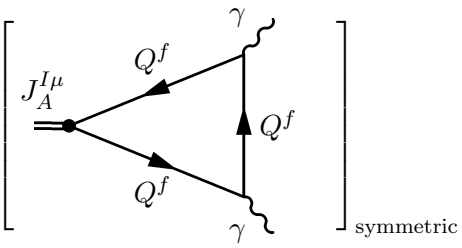


$$\sim \mathrm{tr}\, T^I \ \mathrm{tr}\,\{\tau^a, \tau^b\} = 0 \tag{341}$$

which vanishes because the generators of SU(N_f) are traceless.

From here we would conclude that the nonabelian axial symmetry SU(N_f)$_A$ is nonanomalous. However this is not the whole story since quarks are charged particles that also couple to photons. Hence there is a second potential source of an anomaly coming from the the one-loop triangle diagram coupling $J_A^{I\,\mu}$ to two photons

$$\langle 0|T\left[J_A^{I\,\mu}(x) j^\nu_{\rm em}(x') j^\sigma_{\rm em}(0)\right]|0\rangle = \sum_{f=1}^{N_f}$$



(342)

where $j^\mu_{\rm em}$ is the electromagnetic current

$$j^\mu_{\rm em} = \sum_{f=1}^{N_f} q_f\, \overline{Q}^f \gamma^\mu Q^f, \tag{343}$$

with q_f the electric charge of the f-th quark flavor. A calculation of the diagram in (342) shows the existence of an Adler-Bell-Jackiw anomaly given by

$$\partial_\mu J_A^{I\,\mu} = -\frac{N_c}{16\pi^2}\left[\sum_{f=1}^{N_f}(T^I)_{ff}\, q_f^2\right]\varepsilon^{\mu\nu\sigma\lambda} F_{\mu\nu} F_{\sigma\lambda}, \tag{344}$$

where $F_{\mu\nu}$ is the field strength of the electromagnetic field coupling to the quarks. The only chance for the anomaly to cancel is that the factor between brackets in this equation be identically zero.

[16]The normalization of the generators T^I of the global SU(N_f) is given by $\mathrm{tr}\,(T^I T^J) = \frac{1}{2}\delta^{IJ}$.

Before proceeding let us summarize the results found so far. Because of the presence of anomalies the axial part of the global chiral symmetry, $\mathrm{SU}(N_f)_A$ and $\mathrm{U}(1)_A$ are not realized quantum mechanically in general. We found that $\mathrm{U}(1)_A$ is always affected by an anomaly. However, because the right-hand side of the anomaly equation (340) is a total derivative, the anomalous character of J_A^μ does not explain the absence of $\mathrm{U}(1)_A$ multiplets in the hadron spectrum, since a new current can be constructed which is conserved. In addition, the nonexistence of candidates for a Goldstone boson associated with the right quantum numbers indicates that $\mathrm{U}(1)_A$ is not spontaneously broken either, so it has be explicitly broken somehow. This is the so-called U(1)-problem which was solved by 't Hooft [33], who showed how the contribution of quantum transitions between vacua with topologically nontrivial gauge field configurations (instantons) results in an explicit breaking of this symmetry.

Due to the dynamics of the $\mathrm{SU}(N_c)$ gauge theory the axial nonabelian symmetry is spontaneously broken due to the presence at low energies of a vacuum expectation value for the fermion bilinear $\overline{Q}^f Q^f$

$$\langle 0|\overline{Q}^f Q^f|0\rangle \neq 0 \qquad \text{(No summation in } f\text{!)}. \tag{345}$$

This nonvanishing vacuum expectation value for the quark bilinear actually breaks chiral invariance spontaneously to the vector subgroup $\mathrm{SU}(N_f)_V$, so the only subgroup of the original global symmetry that is realized by the full theory at low energy is

$$\mathrm{U}(N_f)_L \times \mathrm{U}(N_f)_R \longrightarrow \mathrm{SU}(N_f)_V \times \mathrm{U}(1)_B. \tag{346}$$

Associated with this breaking a Goldstone boson should appear with the quantum numbers of the broken nonabelian current. For example, in the case of QCD the Goldstone bosons associated with the spontaneously symmetry breaking induced by the vacuum expectation values $\langle \overline{u}u\rangle$, $\langle \overline{d}d\rangle$ and $\langle (\overline{u}d - \overline{d}u)\rangle$ have been identified as the pions π^0, $\pi^\pm$. These bosons are not exactly massless because of the nonvanishing mass of the u and d quarks. Since the global chiral symmetry is already slightly broken by mass terms in the Lagrangian, the associated Goldstone bosons also have masses although they are very light compared to the masses of other hadrons.

In order to have a better physical understanding of the role of anomalies in the physics of strong interactions we particularize now our analysis of the case of real QCD. Since the u and d quarks are much lighter than the other four flavors, QCD at low energies can be well described by including only these two flavors and ignoring heavier quarks. In this approximation, from our previous discussion we know that the low energy global symmetry of the theory is $\mathrm{SU}(2)_V \times \mathrm{U}(1)_B$, where now the vector group $\mathrm{SU}(2)_V$ is the well-known isospin symmetry. The axial $\mathrm{U}(1)_A$ current is anomalous due to Eq. (340) with $N_f = 2$. In the case of the nonabelian axial symmetry $\mathrm{SU}(2)_A$, taking into account that $q_u = \frac{2}{3}e$ and $q_d = -\frac{1}{3}e$ and that the three generators of SU(2) can be written in terms of the Pauli matrices as $T^K = \frac{1}{2}\sigma^K$ we find

$$\sum_{f=u,d} (T^1)_{ff}\, q_f^2 = \sum_{f=u,d} (T^1)_{ff}\, q_f^2 = 0, \qquad \sum_{f=u,d} (T^3)_{ff}\, q_f^2 = \frac{e^2}{6}. \tag{347}$$

Therefore $J_A^{3\,\mu}$ is anomalous.

Physically, the anomaly in the axial current $J_A^{3\,\mu}$ has an important consequence. In the quark model, the wave function of the neutral pion π^0 is given in terms of those for the u and d quark by

$$|\pi^0\rangle = \frac{1}{\sqrt{2}}\left(|\bar{u}\rangle|u\rangle - |\bar{d}\rangle|d\rangle\right). \tag{348}$$

The isospin quantum numbers of $|\pi^0\rangle$ are those of the generator T^3. Actually the analogy goes further since $\partial_\mu J_A^{3\,\mu}$ is the operator creating a pion π^0 out of the vacuum

$$|\pi^0\rangle \sim \partial_\mu J_A^{3\,\mu}|0\rangle. \tag{349}$$

This leads to the physical interpretation of the triangle diagram (342) with $J_A^{3\,\mu}$ as the one loop contribution to the decay of a neutral pion into two photons

$$\pi^0 \longrightarrow 2\gamma\,. \tag{350}$$

This is an interesting piece of physics. In 1967 Sutherland and Veltman [34] presented a calculation, using current algebra techniques, according to which the decay of the pion into two photons should be suppressed. This however contradicted the experimental evidence that showed the existence of such a decay. The way out to this paradox, as pointed out in [30], is the axial anomaly. What happens is that the current algebra analysis overlooks the ambiguities associated with the regularization of divergences in Quantum Field Theory. A QED evaluation of the triangle diagram leads to a divergent integral that has to be regularized somehow. It is in this process that the Adler-Bell-Jackiw axial anomaly appears resulting in a nonvanishing value for the $\pi^0 \to 2\gamma$ amplitude[17].

The existence of anomalies associated with global currents does not necessarily mean difficulties for the theory. On the contrary, as we saw in the case of the axial anomaly it is its existence what allows for a solution of the Sutherland-Veltman paradox and an explanation of the electromagnetic decay of the pion. The situation, however, is very different if we deal with local symmetries. A quantum mechanical violation of gauge symmetry leads to all kinds of problems, from lack of renormalizability to nondecoupling of negative norm states. This is because the presence of an anomaly in the theory implies that the Gauss' law constraint $\vec{\nabla}\cdot\vec{E}_a = \rho_a$ cannot be consistently implemented in the quantum theory. As a consequence states that classically are eliminated by the gauge symmetry become propagating fields in the quantum theory, thus spoiling the consistency of the theory.

Anomalies in a gauge symmetry can be expected only in chiral theories where left and right-handed fermions transform in different representations of the gauge group. Physically, the most interesting example of such theories is the electroweak sector of the Standard Model where, for example, left handed fermions transform as doublets under SU(2) whereas right-handed fermions are singlets. On the other hand, QCD is free of gauge anomalies since both left- and right-handed quarks transform in the fundamental representation of SU(3).

We consider the Lagrangian

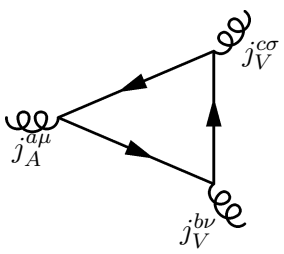
$$\mathcal{L} = -\frac{1}{4}F^{a\,\mu\nu}F^a_{\mu\nu} + i\sum_{i=1}^{N_+}\overline{\psi}^i_+ \not{D}^{\,(+)}\psi^i_+ + i\sum_{j=1}^{N_-}\overline{\psi}^j_- \not{D}^{\,(-)}\psi^j_-, \tag{351}$$

where the chiral fermions $\psi^i_\pm$ transform according to the representations $\tau^a_{i,\pm}$ of the gauge group G ($a = 1,\ldots,\dim G$). The covariant derivatives $D^{(\pm)}_\mu$ are then defined by

$$D^{(\pm)}_\mu\psi^i_\pm = \partial_\mu\psi^i_\pm + igA^K_\mu\tau^K_{i,\pm}\psi^i_\pm. \tag{352}$$

As for global symmetries, anomalies in the gauge symmetry appear in the triangle diagram with one axial and two vector gauge current vertices

$$\langle 0|T\left[j_A^{a\,\mu}(x)j_V^{b\,\nu}(x')j_V^{c\,\sigma}(0)\right]|0\rangle = \Bigg[$$



$$\Bigg]_{\text{symmetric}} \tag{353}$$

[17] An early computation of the triangle diagram for the electromagnetic decay of the pion was made by Steinberger in [31].

where gauge vector and axial currents $j_V^{a\,\mu}$, $j_A^{a\,\mu}$ are given by

$$\begin{aligned} j_V^{a\mu} &= \sum_{i=1}^{N_+} \overline{\psi}_+^i \tau_+^a \gamma^\mu \psi_+^i + \sum_{j=1}^{N_-} \overline{\psi}_-^j \tau_-^a \gamma^\mu \psi_-^j, \\ j_A^{a\mu} &= \sum_{i=1}^{N_+} \overline{\psi}_+^i \tau_+^a \gamma^\mu \psi_+^i - \sum_{i=1}^{N_-} \overline{\psi}_-^j \tau_-^a \gamma^\mu \psi_-^j. \end{aligned} \tag{354}$$

Luckily, we do not have to compute the whole diagram in order to find an anomaly cancellation condition, it is enough if we calculate the overall group theoretical factor. In the case of the diagram in Eq. (353) for every fermion species running in the loop this factor is equal to

$$\mathrm{tr}\,\left[\tau_{i,\pm}^a \{\tau_{i,\pm}^b, \tau_{i,\pm}^c\}\right], \tag{355}$$

where the sign $\pm$ corresponds respectively to the generators of the representation of the gauge group for the left and right-handed fermions. Hence the anomaly cancellation condition reads

$$\sum_{i=1}^{N_+} \mathrm{tr}\,\left[\tau_{i,+}^a \{\tau_{i,+}^b, \tau_{i,+}^c\}\right] - \sum_{j=1}^{N_-} \mathrm{tr}\,\left[\tau_{j,-}^a \{\tau_{j,-}^b, \tau_{j,-}^c\}\right] = 0. \tag{356}$$

Knowing this we can proceed to check the anomaly cancellation in the Standard Model SU(3)×SU(2)×U(1). Left handed fermions (both leptons and quarks) transform as doublets with respect to the SU(2) factor whereas the right-handed components are singlets. The charge with respect to the U(1) part, the hypercharge Y, is determined by the Gell-Mann-Nishijima formula

$$Q = T_3 + Y, \tag{357}$$

where Q is the electric charge of the corresponding particle and T_3 is the eigenvalue with respect to the third generator of the SU(2) group in the corresponding representation: $T_3 = \frac{1}{2}\sigma^3$ for the doublets and $T_3 = 0$ for the singlets. For the first family of quarks (u, d) and leptons (e, ν_e) we have the following field content

$$\begin{array}{lccc} \text{quarks:} & \begin{pmatrix} u^\alpha \\ d^\alpha \end{pmatrix}_{L,\frac{1}{6}} & u^\alpha_{R,\frac{2}{3}} & d^\alpha_{R,\frac{2}{3}} \\ \text{leptons:} & \begin{pmatrix} \nu_e \\ e \end{pmatrix}_{L,-\frac{1}{2}} & e_{R,-1} & \end{array} \tag{358}$$

where $\alpha = 1, 2, 3$ labels the color quantum number and the subscript indicates the value of the weak hypercharge Y. Denoting the representations of SU(3)×SU(2)×U(1) by $(n_c, n_w)_Y$, with n_c and n_w the representations of SU(3) and SU(2) respectively and Y the hypercharge, the matter content of the Standard Model consists of a three family replication of the representations:

$$\begin{array}{lccc} \text{left-handed fermions:} & (3,2)^L_{\frac{1}{6}} & (1,2)^L_{-\frac{1}{2}} & \\ \text{right-handed fermions:} & (3,1)^R_{\frac{2}{3}} & (3,1)^R_{-\frac{1}{3}} & (1,1)^R_{-1}. \end{array} \tag{359}$$

In computing the triangle diagram we have 10 possibilities depending on which factor of the gauge group

SU(3)×SU(2)×U(1) couples to each vertex:

$\mathrm{SU}(3)^3$	$\mathrm{SU}(2)^3$	$\mathrm{U}(1)^3$
$\mathrm{SU}(3)^2\,\mathrm{SU}(2)$	$\mathrm{SU}(2)^2\,\mathrm{U}(1)$	
$\mathrm{SU}(3)^2\,\mathrm{U}(1)$	$\mathrm{SU}(2)\,\mathrm{U}(1)^2$	
$\mathrm{SU}(3)\,\mathrm{SU}(2)^2$		
$\mathrm{SU}(3)\,\mathrm{SU}(2)\,\mathrm{U}(1)$		
$\mathrm{SU}(3)\,\mathrm{U}(1)^2$		

It is easy to check that some of them do not give rise to anomalies. For example the anomaly for the $\mathrm{SU}(3)^3$ case cancels because left and right-handed quarks transform in the same representation. In the case of $\mathrm{SU}(2)^3$ the cancellation happens term by term because of the Pauli matrices identity $\sigma^a\sigma^b = \delta^{ab} + i\varepsilon^{abc}\sigma^c$ that leads to

$$\mathrm{tr}\left[\sigma^a\{\sigma^b,\sigma^c\}\right] = 2\,(\mathrm{tr}\,\sigma^a)\,\delta^{bc} = 0. \tag{360}$$

However the hardest anomaly cancellation condition to satisfy is the one with three U(1)'s. In this case the absence of anomalies within a single family is guaranteed by the nontrivial identity

$$\begin{aligned}\sum_{\text{left}} Y_+^3 - \sum_{\text{right}} Y_-^3 &= 3\times 2\times\left(\frac{1}{6}\right)^3 + 2\times\left(-\frac{1}{2}\right)^3 - 3\times\left(\frac{2}{3}\right)^3 - 3\times\left(-\frac{1}{3}\right)^3 - (-1)^3 \\ &= \left(-\frac{3}{4}\right) + \left(\frac{3}{4}\right) = 0.\end{aligned} \tag{361}$$

It is remarkable that the anomaly exactly cancels between leptons and quarks. Notice that this result holds even if a right-handed sterile neutrino is added since such a particle is a singlet under the whole Standard Model gauge group and therefore does not contribute to the triangle diagram. Therefore we see how the matter content of the Standard Model conspires to yield a consistent quantum field theory.

In all our discussion of anomalies we only considered the computation of one-loop diagrams. It may happen that higher loop orders impose additional conditions. Fortunately this is not so: the Adler-Bardeen theorem [35] guarantees that the axial anomaly only receives contributions from one loop diagrams. Therefore, once anomalies are canceled (if possible) at one loop we know that there will be no new conditions coming from higher-loop diagrams in perturbation theory.

The Adler-Bardeen theorem, however, only applies in perturbation theory. It is nonetheless possible that nonperturbative effects can result in the quantum violation of a gauge symmetry. This is precisely the case pointed out by Witten [36] with respect to the SU(2) gauge symmetry of the Standard Model. In this case the problem lies in the nontrivial topology of the gauge group SU(2). The invariance of the theory with respect to gauge transformations which are not in the connected component of the identity makes all correlation functions equal to zero. Only when the number of left-handed SU(2) fermion doublets is even gauge invariance allows for a nontrivial theory. It is again remarkable that the family structure of the Standard Model makes this anomaly to cancel

$$3\times\begin{pmatrix} u \\ d \end{pmatrix}_L + 1\times\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L = 4\ \text{SU(2)-doublets}, \tag{362}$$

where the factor of 3 comes from the number of colors.

8 Renormalization

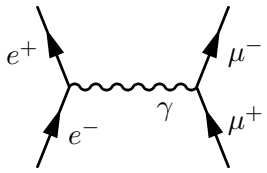
8.1 Removing infinities

From its very early stages, Quantum Field Theory was faced with infinities. They emerged in the calculation of most physical quantities, such as the correction to the charge of the electron due to the interactions with the radiation field. The way these divergences where handled in the 1940s, starting with Kramers, was physically very much in the spirit of the Quantum Theory emphasis in observable quantities: since the observed magnitude of physical quantities (such as the charge of the electron) is finite, this number should arise from the addition of a "bare" (unobservable) value and the quantum corrections. The fact that both of these quantities were divergent was not a problem physically, since only its finite sum was an observable quantity. To make thing mathematically sound, the handling of infinities requires the introduction of some regularization procedure which cuts the divergent integrals off at some momentum scale Λ. Morally speaking, the physical value of an observable $\mathcal{O}_{\rm physical}$ is given by

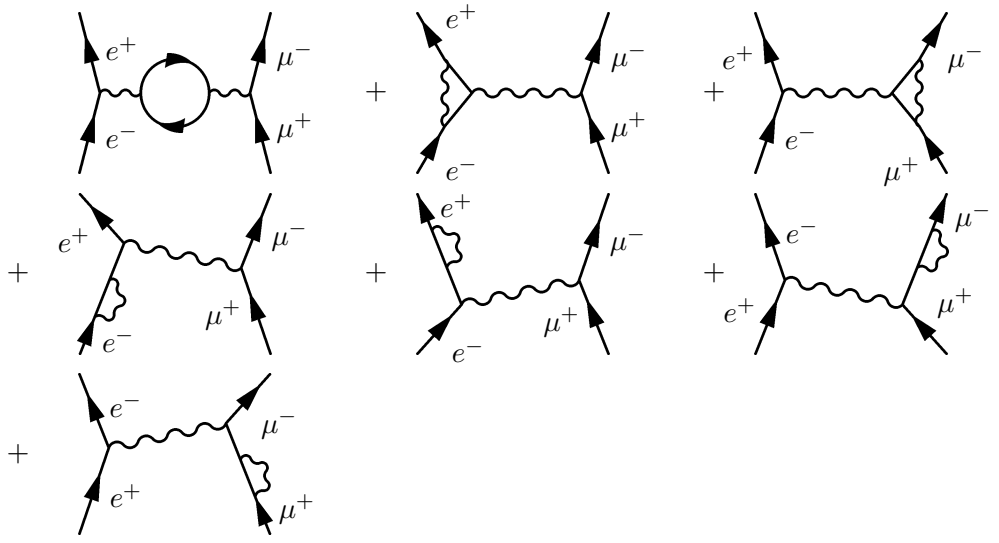
$$\mathcal{O}_{\rm physical} = \lim_{\Lambda\to\infty} \left[\mathcal{O}(\Lambda)_{\rm bare} + \Delta\mathcal{O}(\Lambda)_\hbar\right], \tag{363}$$

where $\Delta\mathcal{O}(\Lambda)_\hbar$ represents the regularized quantum corrections.

To make this qualitative discussion more precise we compute the corrections to the electric charge in Quantum Electrodynamics. We consider the process of annihilation of an electron-positron pair to create a muon-antimuon pair $e^-e^+ \to \mu^+\mu^-$. To lowest order in the electric charge e the only diagram contributing is

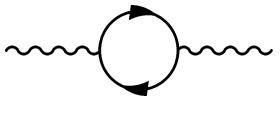


However, the corrections at order e^4 to this result requires the calculation of seven more diagrams



In order to compute the renormalization of the charge we consider the first diagram which takes into account the first correction to the propagator of the virtual photon interchanged between the pairs

due to vacuum polarization. We begin by evaluating



$$= \frac{-i\eta^{\mu\alpha}}{q^2+i\epsilon}\left[\alpha \quad \beta\right]\frac{-i\eta^{\beta\nu}}{q^2+i\epsilon}, \tag{364}$$

where the diagram between brackets is given by

$$\equiv \Pi^{\alpha\beta}(q) = i^2(-ie)^2(-1)\int \frac{d^4k}{(2\pi)^4}\frac{\mathrm{Tr}\,(\not{k}+m_e)\gamma^\alpha(\not{k}+\not{q}+m_e)\gamma^\beta}{[k^2-m_e^2+i\epsilon]\,[(k+q)^2-m_e^2+i\epsilon]}. \tag{365}$$

Physically this diagram includes the correction to the propagator due to the polarization of the vacuum, i.e. the creation of virtual electron-positron pairs by the propagating photon. The momentum q is the total momentum of the electron-positron pair in the intermediate channel.

It is instructive to look at this diagram from the point of view of perturbation theory in nonrelativistic Quantum Mechanics. In each vertex the interaction consists of the annihilation (resp. creation) of a photon and the creation (resp. annihilation) of an electron-positron pair. This can be implemented by the interaction Hamiltonian

$$H_{\rm int} = e\int d^3x\,\overline{\psi}\gamma^\mu\psi A_\mu. \tag{366}$$

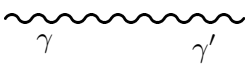
All fields inside the integral can be expressed in terms of the corresponding creation-annihilation operators for photons, electrons and positrons. In Quantum Mechanics, the change in the wave function at first order in the perturbation $H_{\rm int}$ is given by

$$|\gamma,{\rm in}\rangle = |\gamma,{\rm in}\rangle_0 + \sum_n \frac{\langle n|H_{\rm int}|\gamma,{\rm in}\rangle_0}{E_{\rm in}-E_n}|n\rangle \tag{367}$$

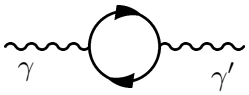
and similarly for $|\gamma,{\rm out}\rangle$, where we have denoted symbolically by $|n\rangle$ all the possible states of the electron-positron pair. Since these states are orthogonal to $|\gamma,{\rm in}\rangle_0$, $|\gamma,{\rm out}\rangle_0$, we find torder e^2

$$\langle\gamma,{\rm in}|\gamma',{\rm out}\rangle = {}_0\langle\gamma,{\rm in}|\gamma',{\rm out}\rangle_0 + \sum_n \frac{{}_0\langle\gamma,{\rm in}|H_{\rm int}|n\rangle\,\langle n|H_{\rm int}|\gamma',{\rm out}\rangle_0}{(E_{\rm in}-E_n)(E_{\rm out}-E_n)} + \mathcal{O}(e^4). \tag{368}$$

Hence, we see that the diagram of Eq. (364) really corresponds to the order-e^2 correction to the photon propagator $\langle\gamma,{\rm in}|\gamma',{\rm out}\rangle$



$$\longrightarrow \quad {}_0\langle\gamma,{\rm in}|\gamma',{\rm out}\rangle_0$$



$$\longrightarrow \quad \sum_n \frac{\langle\gamma,{\rm in}|H_{\rm int}|n\rangle\,\langle n|H_{\rm int}|\gamma',{\rm out}\rangle}{(E_{\rm in}-E_n)(E_{\rm out}-E_n)}. \tag{369}$$

Once we understood the physical meaning of the Feynman diagram to be computed we proceed to its evaluation. In principle there is no problem in computing the integral in Eq. (364) for nonzero values of the electron mass. However since here we are going to be mostly interested in seeing how the divergence of the integral results in a scale-dependent renormalization of the electric charge, we will set $m_e = 0$. This is something safe to do, since in the case of this diagram we are not inducing new infrared divergences in taking the electron as massless. Implementing gauge invariance and using standard techniques in the computation of Feynman diagrams (see references [1]- [11]) the polarization tensor $\Pi_{\mu\nu}(q)$ defined in Eq. (365) can be written as

$$\Pi_{\mu\nu}(q) = \left(q^2\eta_{\mu\nu} - q_\mu q_\nu\right)\Pi(q^2) \tag{370}$$

with

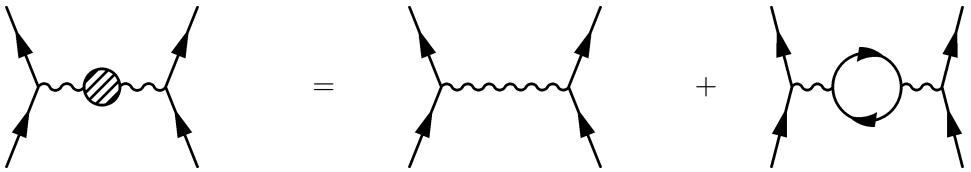
$$\Pi(q) = 8e^2\int_0^1 dx \int \frac{d^4k}{(2\pi)^4}\frac{x(1-x)}{[k^2 - m^2 + x(1-x)q^2 + i\epsilon]^2} \tag{371}$$

To handle this divergent integral we have to figure out some procedure to render it finite. This can be done in several ways, but here we choose to cut the integrals off at a high energy scale Λ, where new physics might be at work, $|p| < \Lambda$. This gives the result

$$\Pi(q^2) \simeq \frac{e^2}{12\pi^2}\log\left(\frac{q^2}{\Lambda^2}\right) + \text{finite terms.} \tag{372}$$

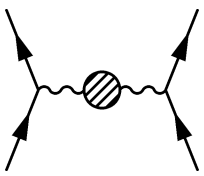
If we would send the cutoff to infinity $\Lambda \to \infty$ the divergence blows up and something has to be done about it.

If we want to make sense out of this, we have to go back to the physical question that led us to compute Eq. (364). Our primordial motivation was to compute the corrections to the annihilation of two electrons into two muons. Including the correction to the propagator of the virtual photon we have



$$\begin{aligned} &= \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\frac{e^2}{4\pi q^2}\left(\overline{v}_\mu\gamma^\beta u_\mu\right) + \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\frac{e^2}{4\pi q^2}\Pi(q^2)\left(\overline{v}_\mu\gamma^\beta u_\mu\right) \\ &= \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\left\{\frac{e^2}{4\pi q^2}\left[1 + \frac{e^2}{12\pi^2}\log\left(\frac{q^2}{\Lambda^2}\right)\right]\right\}\left(\overline{v}_\mu\gamma^\beta u_\mu\right). \end{aligned} \tag{373}$$

Now let us imagine that we are performing a $e^-\,e^+ \to \mu^-\mu^+$ with a center of mass energy μ. From the previous result we can identify the effective charge of the particles at this energy scale $e(\mu)$ as



$$= \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\left[\frac{e(\mu)^2}{4\pi q^2}\right]\left(\overline{v}_\mu\gamma^\beta u_\mu\right). \tag{374}$$

This charge, $e(\mu)$, is the quantity that is physically measurable in our experiment. Now we can make sense of the formally divergent result (373) by assuming that the charge appearing in the classical Lagrangian of QED is just a "bare" value that depends on the scale Λ at which we cut off the theory, $e \equiv e(\Lambda)_{\text{bare}}$. In order to reconcile (373) with the physical results (374) we must assume that the dependence of the bare (unobservable) charge $e(\Lambda)_{\text{bare}}$ on the cutoff Λ is determined by the identity

$$e(\mu)^2 = e(\Lambda)^2_{\text{bare}}\left[1 + \frac{e(\Lambda)^2_{\text{bare}}}{12\pi^2}\log\left(\frac{\mu^2}{\Lambda^2}\right)\right]. \tag{375}$$

If we still insist in removing the cutoff, $\Lambda \to \infty$ we have to send the bare charge to zero $e(\Lambda)_{\rm bare} \to 0$ in such a way that the effective coupling has the finite value given by the experiment at the energy scale μ. It is not a problem, however, that the bare charge is small for large values of the cutoff, since the only measurable quantity is the effective charge that remains finite. Therefore all observable quantities should be expressed in perturbation theory as a power series in the physical coupling $e(\mu)^2$ and not in the unphysical bare coupling $e(\Lambda)_{\rm bare}$.

8.2 The beta-function and asymptotic freedom

We can look at the previous discussion, an in particular Eq. (375), from a different point of view. In order to remove the ambiguities associated with infinities we have been forced to introduce a dependence of the coupling constant on the energy scale at which a process takes place. From the expression of the physical coupling in terms of the bare charge (375) we can actually eliminate the cutoff Λ, whose value after all should not affect the value of physical quantities. Taking into account that we are working in perturbation theory in $e(\mu)^2$, we can express the bare charge $e(\Lambda)^2_{\rm bare}$ in terms of $e(\mu)^2$ as

$$e(\Lambda)^2_{\rm bare} = e(\mu)^2 \left[1 - \frac{e(\mu)^2}{12\pi^2} \log\left(\frac{\mu^2}{\Lambda^2}\right)\right] + \mathcal{O}[e(\mu)^6]. \tag{376}$$

This expression allow us to eliminate all dependence in the cutoff in the expression of the effective charge at a scale μ by replacing $e(\Lambda)_{\rm bare}$ in Eq. (375) by the one computed using (376) at a given reference energy scale μ_0

$$e(\mu)^2 = e(\mu_0)^2 \left[1 + \frac{e(\mu_0)^2}{12\pi^2} \log\left(\frac{\mu^2}{\mu_0^2}\right)\right]. \tag{377}$$

From this equation we can compute, at this order in perturbation theory, the effective value of the coupling constant at an energy μ, once we know its value at some reference energy scale μ_0. In the case of the electron charge we can use as a reference Thompson's scattering at energies of the order of the electron mass $m_e \simeq 0.5$ MeV, at where the value of the electron charge is given by the well known value

$$e(m_e)^2 \simeq \frac{1}{137}. \tag{378}$$

With this we can compute $e(\mu)^2$ at any other energy scale applying Eq. (377), for example at the electron mass $\mu = m_e \simeq 0.5\,$MeV. However, in computing the electromagnetic coupling constant at any other scale we must take into account the fact that other charged particles can run in the loop in Eq. (373). Suppose, for example, that we want to calculate the fine structure constant at the mass of the Z^0-boson $\mu = M_Z \equiv 92$ GeV. Then we should include in Eq. (377) the effect of other fermionic Standard Model fields with masses below M_Z. Doing this, we find[18]

$$e(M_Z)^2 = e(m_e)^2 \left[1 + \frac{e(m_e)^2}{12\pi^2} \left(\sum_i q_i^2\right) \log\left(\frac{M_Z^2}{m_e^2}\right)\right], \tag{379}$$

where q_i is the charge in units of the electron charge of the i-th fermionic species running in the loop and we sum over all fermions with masses below the mass of the Z^0 boson. This expression shows how the electromagnetic coupling grows with energy. However, in order to compare with the experimental value of $e(M_Z)^2$ it is not enough with including the effect of fermionic fields, since also the $W^\pm$ bosons

[18] In the first version of these notes the argument used to show the growing of the electromagnetic coupling constant could have led to confusion to some readers. To avoid this potential problem we include in the equation for the running coupling $e(\mu)^2$ the contribution of all fermions with masses below M_Z. We thank Lubos Motl for bringing this issue to our attention.

can run in the loop ($M_W < M_Z$). Taking this into account, as well as threshold effects, the value of the electron charge at the scale M_Z is found to be [37]

$$e(M_Z)^2 \simeq \frac{1}{128.9} \ . \tag{380}$$

This growing of the effective fine structure constant with energy can be understood heuristically by remembering that the effect of the polarization of the vacuum shown in the diagram of Eq. (364) amounts to the creation of a plethora of electron-positron pairs around the location of the charge. These virtual pairs behave as dipoles that, as in a dielectric medium, tend to screen this charge and decreasing its value at long distances (i.e. lower energies).

The variation of the coupling constant with energy is usually encoded in Quantum Field Theory in the *beta function* defined by

$$\beta(g) = \mu \frac{dg}{d\mu}. \tag{381}$$

In the case of QED the beta function can be computed from Eq. (377) with the result

$$\beta(e)_{\rm QED} = \frac{e^3}{12\pi^2}. \tag{382}$$

The fact that the coefficient of the leading term in the beta-function is positive $\beta_0 \equiv \frac{1}{6\pi} > 0$ gives us the overall behavior of the coupling as we change the scale. Eq. (382) means that, if we start at an energy where the electric coupling is small enough for our perturbative treatment to be valid, the effective charge grows with the energy scale. This growing of the effective coupling constant with energy means that QED is infrared safe, since the perturbative approximation gives better and better results as we go to lower energies. Actually, because the electron is the lighter electrically charged particle and has a finite nonvanishing mass the running of the fine structure constant stops at the scale m_e in the well-known value $\frac{1}{137}$. Would other charged fermions with masses below m_e be present in Nature, the effective value of the fine structure constant in the interaction between these particles would run further to lower values at energies below the electron mass.

On the other hand if we increase the energy scale $e(\mu)^2$ grows until at some scale the coupling is of order one and the perturbative approximation breaks down. In QED this is known as the problem of the Landau pole but in fact it does not pose any serious threat to the reliability of QED perturbation theory: a simple calculation shows that the energy scale at which the theory would become strongly coupled is $\Lambda_{\rm Landau} \simeq 10^{277}$ GeV. However, we know that QED does not live that long! At much lower scales we expect electromagnetism to be unified with other interactions, and even if this is not the case we will enter the uncharted territory of quantum gravity at energies of the order of 10^{19} GeV.

So much for QED. The next question that one may ask at this stage is whether it is possible to find quantum field theories with a behavior opposite to that of QED, i.e. such that they become weakly coupled at high energies. This is not a purely academic question. In the late 1960s a series of deep-inelastic scattering experiments carried out at SLAC showed that the quarks behave essentially as free particles inside hadrons. The apparent problem was that no theory was known at that time that would become free at very short distances: the example set by QED seem to be followed by all the theories that were studied. This posed a very serious problem for Quantum Field Theory as a way to describe subnuclear physics, since it seemed that its predictive power was restricted to electrodynamics but failed miserably when applied to describe strong interactions.

Nevertheless, this critical time for Quantum Field Theory turned out to be its finest hour. In 1973 David Gross and Frank Wilczek [38] and David Politzer [39] showed that nonabelian gauge theories can actually display the required behavior. For the QCD Lagrangian in Eq. (331) the beta function is given

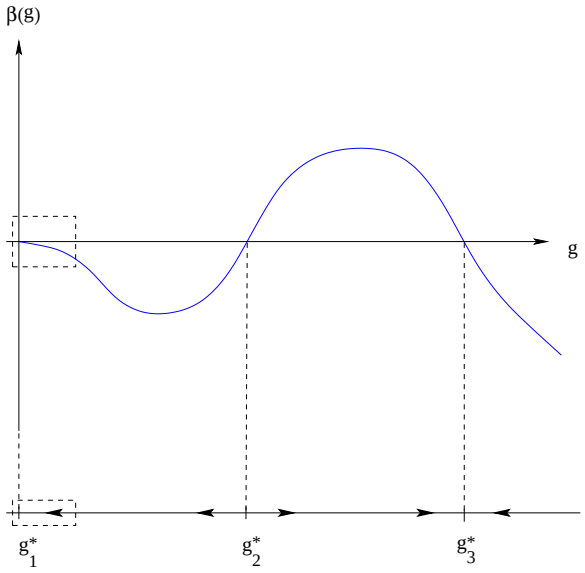


Fig. 14: Beta function for a hypothetical theory with three fixed points g_1^*, g_2^* and g_3^*. A perturbative analysis would capture only the regions shown in the boxes.

by[19]

$$\beta(g) = -\frac{g^3}{16\pi^2}\left[\frac{11}{3}N_c - \frac{2}{3}N_f\right]. \tag{383}$$

In particular, for real QCD ($N_C = 3$, $N_f = 6$) we have that $\beta(g) = -\frac{7g^3}{16\pi^2} < 0$. This means that for a theory that is weakly coupled at an energy scale μ_0 the coupling constant decreases as the energy increases $\mu \to \infty$. This explain the apparent freedom of quarks inside the hadrons: when the quarks are very close together their effective color charge tend to zero. This phenomenon is called *asymptotic freedom*.

Asymptotic free theories display a behavior that is opposite to that found above in QED. At high energies their coupling constant approaches zero whereas at low energies they become strongly coupled (infrared slavery). This features are at the heart of the success of QCD as a theory of strong interactions, since this is exactly the type of behavior found in quarks: they are quasi-free particles inside the hadrons but the interaction potential potential between them increases at large distances.

Although asymptotic free theories can be handled in the ultraviolet, they become extremely complicated in the infrared. In the case of QCD it is still to be understood (at least analytically) how the theory confines color charges and generates the spectrum of hadrons, as well as the breaking of the chiral symmetry (345).

In general, the ultraviolet and infrared properties of a theory are controlled by the fixed points of the beta function, i.e. those values of the coupling constant g for which it vanishes

$$\beta(g^*) = 0. \tag{384}$$

Using perturbation theory we have seen that for both QED and QCD one of such fixed points occurs at zero coupling, $g^* = 0$. However, our analysis also showed that the two theories present radically different behavior at high and low energies. From the point of view of the beta function, the difference lies in the energy regime at which the coupling constant approaches its critical value. This is in fact governed by the sign of the beta function around the critical coupling.

[19]The expression of the beta function of QCD was also known to 't Hooft [40]. There are even earlier computations in the russian literature [41].

We have seen above that when the beta function is negative close to the fixed point (the case of QCD) the coupling tends to its critical value, $g^* = 0$, as the energy is increased. This means that the critical point is *ultraviolet stable*, i.e. it is an attractor as we evolve towards higher energies. If, on the contrary, the beta function is positive (as it happens in QED) the coupling constant approaches the critical value as the energy decreases. This is the case of an *infrared stable* fixed point.

This analysis that we have motivated with the examples of QED and QCD is completely general and can be carried out for any quantum field theory. In Fig. 14 we have represented the beta function for a hypothetical theory with three fixed points located at couplings g_1^*, g_2^* and g_3^*. The arrows in the line below the plot represent the evolution of the coupling constant as the energy increases. From the analysis presented above we see that $g_1^* = 0$ and g_3^* are ultraviolet stable fixed points, while the fixed point g_2^* is infrared stable.

In order to understand the high and low energy behavior of a quantum field theory it is then crucial to know the structure of the beta functions associated with its couplings. This can be a very difficult task, since perturbation theory only allows the study of the theory around "trivial" fixed points, i.e. those that occur at zero coupling like the case of g_1^* in Fig. 14. On the other hand, any "nontrivial" fixed point occurring in a theory (like g_2^* and g_3^*) cannot be captured in perturbation theory and requires a full nonperturbative analysis.

The moral to be learned from our discussion above is that dealing with the ultraviolet divergences in a quantum field theory has the consequence, among others, of introducing an energy dependence in the measured value of the coupling constants of the theory (for example the electric charge in QED). This happens even in the case of renormalizable theories without mass terms. These theories are scale invariant at the classical level because the action does not contain any dimensionful parameter. In this case the running of the coupling constants can be seen as resulting from a quantum breaking of classical scale invariance: different energy scales in the theory are distinguished by different values of the coupling constants. Remembering what we learned in Section 7, we conclude that classical scale invariance is an anomalous symmetry. One heuristic way to see how the conformal anomaly comes about is to notice that the regularization of an otherwise scale invariant field theory requires the introduction of an energy scale (e.g. a cutoff). This breaking of scale invariance cannot be restored after renormalization.

Nevertheless, scale invariance is not lost forever in the quantum theory. It is recovered at the fixed points of the beta function where, by definition, the coupling does not run. To understand how this happens we go back to a scale invariant classical field theory whose field $\phi(x)$ transform under coordinate rescalings as

$$x^\mu \longrightarrow \lambda x^\mu, \qquad \phi(x) \longrightarrow \lambda^{-\Delta}\phi(\lambda^{-1}x), \tag{385}$$

where Δ is called the canonical scaling dimension of the field. An example of such a theory is a massless ϕ^4 theory in four dimensions

$$\mathcal{L} = \frac{1}{2}\partial_\mu\phi\,\partial^\mu\phi - \frac{g}{4!}\phi^4, \tag{386}$$

where the scalar field has canonical scaling dimension $\Delta = 1$. The Lagrangian density transforms as

$$\mathcal{L} \longrightarrow \lambda^{-4}\mathcal{L}[\phi] \tag{387}$$

and the classical action remains invariant[20].

If scale invariance is preserved under quantization, the Green's functions transform as

$$\langle\Omega|T[\phi'(x_1)\dots\phi'(x_n)]|\Omega\rangle = \lambda^{n\Lambda}\langle\Omega|T[\phi(\lambda^{-1}x_1)\dots\phi(\lambda^{-1}x_n)]|\Omega\rangle. \tag{388}$$

[20]In a D-dimensional theory the canonical scaling dimensions of the fields coincide with its engineering dimension: $\Delta = \frac{D-2}{2}$ for bosonic fields and $\Delta = \frac{D-1}{2}$ for fermionic ones. For a Lagrangian with no dimensionful parameters classical scale invariance follows then from dimensional analysis.

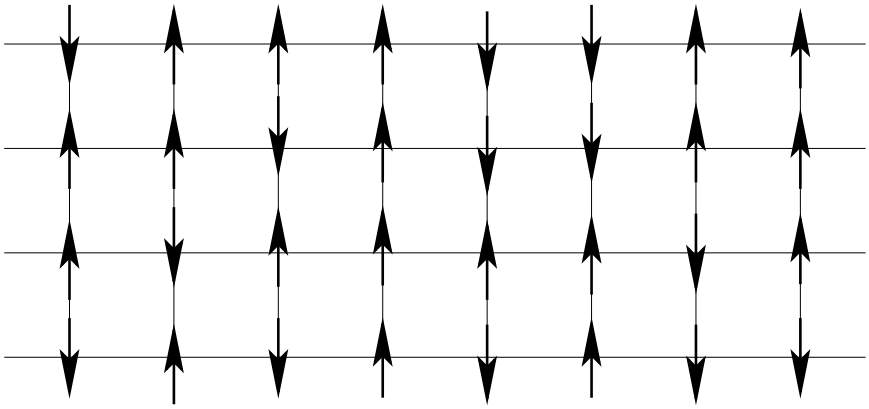


Fig. 15: Systems of spins in a two-dimensional square lattice.

This is precisely what happens in a free theory. In an interacting theory the running of the coupling constant destroys classical scale invariance at the quantum level. Despite of this, at the fixed points of the beta function the Green's functions transform again according to (388) where Δ is replaced by

$$\Delta_{\rm anom} = \Delta + \gamma^*. \tag{389}$$

The canonical scaling dimension of the fields are corrected by γ^*, which is called the anomalous dimension. They carry the dynamical information about the high-energy behavior of the theory.

8.3 The renormalization group

In spite of its successes, the renormalization procedure presented above can be seen as some kind of prescription or recipe to get rid of the divergences in an ordered way. This discomfort about renormalization was expressed in occasions by comparing it with "sweeping the infinities under the rug". However thanks to Ken Wilson to a large extent [42] the process of renormalization is now understood in a very profound way as a procedure to incorporate the effects of physics at high energies by modifying the value of the parameters that appear in the Lagrangian.

Statistical mechanics. Wilson's ideas are both simple and profound and consist in thinking about Quantum Field Theory as the analog of a thermodynamical description of a statistical system. To be more precise, let us consider an Ising spin system in a two-dimensional square lattice as the one depicted in Fig 15. In terms of the spin variables $s_i = \pm\frac{1}{2}$, where i labels the lattice site, the Hamiltonian of the system is given by

$$H = -J \sum_{\langle i,j \rangle} s_i \, s_j, \tag{390}$$

where $\langle i, j \rangle$ indicates that the sum extends over nearest neighbors and J is the coupling constant between neighboring spins (here we consider that there is no external magnetic field). The starting point to study the statistical mechanics of this system is the partition function defined as

$$\mathcal{Z} = \sum_{\{s_i\}} e^{-\beta H}, \tag{391}$$

where the sum is over all possible configurations of the spins and $\beta = \frac{1}{T}$ is the inverse temperature. For $J > 0$ the Ising model presents spontaneous magnetization below a critical temperature T_c, in any dimension higher than one. Away from this temperature correlations between spins decay exponentially at large distances

$$\langle s_i s_j \rangle \sim e^{-\frac{|x_{ij}|}{\xi}}, \tag{392}$$

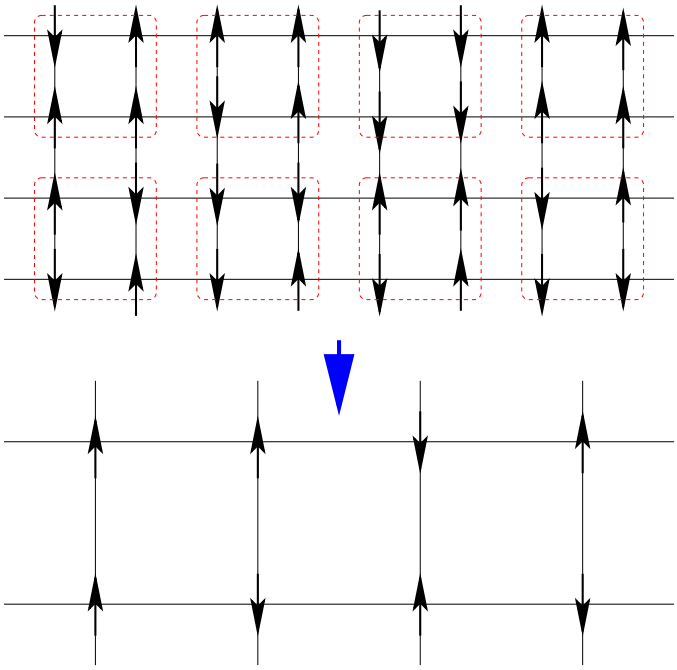


Fig. 16: Decimation of the spin lattice. Each block in the upper lattice is replaced by an effective spin computed according to the rule (394). Notice also that the size of the lattice spacing is doubled in the process.

with $|x_{ij}|$ the distance between the spins located in the i-th and j-th sites of the lattice. This expression serves as a definition of the correlation length ξ which sets the characteristic length scale at which spins can influence each other by their interaction through their nearest neighbors.

Suppose now that we are interested in a macroscopic description of this spin system. We can capture the relevant physics by integrating out somehow the physics at short scales. A way in which this can be done was proposed by Leo Kadanoff [43] and consists in dividing our spin system in spin-blocks like the ones showed in Fig 16. Now we can construct another spin system where each spin-block of the original lattice is replaced by an effective spin calculated according to some rule from the spins contained in each block B_a

$$\{s_i : i \in B_a\} \quad \longrightarrow \quad {s_a}^{(1)}. \tag{393}$$

For example we can define the effective spin associated with the block B_a by taking the majority rule with an additional prescription in case of a draw

$${s_a}^{(1)} = \frac{1}{2}\text{sgn}\left(\sum_{i\in B_a} s_i\right), \tag{394}$$

where we have used the sign function, $\text{sign}(x) \equiv \frac{x}{|x|}$, with the additional definition $\text{sgn}(0) = 1$. This procedure is called decimation and leads to a new spin system with a doubled lattice space.

The idea now is to rewrite the partition function (391) only in terms of the new effective spins ${s_a}^{(1)}$. Then we start by splitting the sum over spin configurations into two nested sums, one over the spin blocks and a second one over the spins within each block

$$\mathcal{Z} = \sum_{\{\vec{s}\}} e^{-\beta H[s_i]} = \sum_{\{\vec{s}^{(1)}\}} \sum_{\{\vec{s}\in B_a\}} \delta\left[{s_a}^{(1)} - \text{sign}\left(\sum_{i\in B_a} s_i\right)\right] e^{-\beta H[s_i]}. \tag{395}$$

The interesting point now is that the sum over spins inside each block can be written as the exponential of a new effective Hamiltonian depending only on the effective spins, $H^{(1)}[s_a{}^{(1)}]$

$$\sum_{\{s\in B_a\}} \delta\left[s_a{}^{(1)} - \text{sign}\left(\sum_{i\in B_a} s_i\right)\right] e^{-\beta H[s_i]} = e^{-\beta H^{(1)}[s_a{}^{(1)}]}. \tag{396}$$

The new Hamiltonian is of course more complicated

$$H^{(1)} = -J^{(1)} \sum_{\langle i,j\rangle} s_i^{(1)} s_j^{(1)} + \dots \tag{397}$$

where the dots stand for other interaction terms between the effective block spins. This new terms appear because in the process of integrating out short distance physics we induce interactions between the new effective degrees of freedom. For example the interaction between the spin block variables $s_i^{(1)}$ will in general not be restricted to nearest neighbors in the new lattice. The important point is that we have managed to rewrite the partition function solely in terms of this new (renormalized) spin variables $s^{(1)}$ interacting through a new Hamiltonian $H^{(1)}$

$$\mathcal{Z} = \sum_{\{s^{(1)}\}} e^{-\beta H^{(1)}[s_a{}^{(1)}]}. \tag{398}$$

Let us now think about the space of all possible Hamiltonians for our statistical system including all kinds of possible couplings between the individual spins compatible with the symmetries of the system. If denote by $\mathcal{R}$ the decimation operation, our previous analysis shows that $\mathcal{R}$ defines a map in this space of Hamiltonians

$$\mathcal{R} : H \to H^{(1)}. \tag{399}$$

At the same time the operation $\mathcal{R}$ replaces a lattice with spacing a by another one with double spacing $2a$. As a consequence the correlation length in the new lattice measured in units of the lattice spacing is divided by two, $\mathcal{R} : \xi \to \frac{\xi}{2}$.

Now we can iterate the operation $\mathcal{R}$ an indefinite number of times. Eventually we might reach a Hamiltonian $H_\star$ that is not further modified by the operation $\mathcal{R}$

$$H \xrightarrow{\mathcal{R}} H^{(1)} \xrightarrow{\mathcal{R}} H^{(2)} \xrightarrow{\mathcal{R}} \dots \xrightarrow{\mathcal{R}} H_\star. \tag{400}$$

The fixed point Hamiltonian $H_\star$ is *scale invariant* because it does not change as $\mathcal{R}$ is performed. Notice that because of this invariance the correlation length of the system at the fixed point do not change under $\mathcal{R}$. This fact is compatible with the transformation $\xi \to \frac{\xi}{2}$ only if $\xi = 0$ or $\xi = \infty$. Here we will focus in the case of nontrivial fixed points with infinite correlation length.

The space of Hamiltonians can be parametrized by specifying the values of the coupling constants associated with all possible interaction terms between individual spins of the lattice. If we denote by $\mathcal{O}_a[s_i]$ these (possibly infinite) interaction terms, the most general Hamiltonian for the spin system under study can be written as

$$H[s_i] = \sum_{a=1}^{\infty} \lambda_a \mathcal{O}_a[s_i], \tag{401}$$

where $\lambda_a \in \mathbb{R}$ are the coupling constants for the corresponding operators. These constants can be thought of as coordinates in the space of all Hamiltonians. Therefore the operation $\mathcal{R}$ defines a transformation in the set of coupling constants

$$\mathcal{R} : \lambda_a \longrightarrow \lambda_a^{(1)}. \tag{402}$$

For example, in our case we started with a Hamiltonian in which only one of the coupling constants is different from zero (say $\lambda_1 = -J$). As a result of the decimation $\lambda_1 \equiv -J \to -J^{(1)}$ while some of the originally vanishing coupling constants will take a nonzero value. Of course, for the fixed point Hamiltonian the coupling constants do not change under the scale transformation $\mathcal{R}$.

Physically the transformation $\mathcal{R}$ integrates out short distance physics. The consequence for physics at long distances is that we have to replace our Hamiltonian by a new one with different values for the coupling constants. That is, our ignorance of the details of the physics going on at short distances result in a *renormalization* of the coupling constants of the Hamiltonian that describes the long range physical processes. It is important to stress that although $\mathcal{R}$ is sometimes called a renormalization group transformation in fact this is a misnomer. Transformations between Hamiltonians defined by $\mathcal{R}$ do not form a group: since these transformations proceed by integrating out degrees of freedom at short scales they cannot be inverted.

In statistical mechanics fixed points under renormalization group transformations with $\xi = \infty$ are associated with phase transitions. From our previous discussion we can conclude that the space of Hamiltonians is divided in regions corresponding to the basins of attraction of the different fixed points. We can ask ourselves now about the stability of those fixed points. Suppose we have a statistical system described by a fixed-point Hamiltonian $H_\star$ and we perturb it by changing the coupling constant associated with an interaction term $\mathcal{O}$. This is equivalent to replace $H_\star$ by the perturbed Hamiltonian

$$H = H_\star + \delta\lambda\, \mathcal{O}, \tag{403}$$

where $\delta\lambda$ is the perturbation of the coupling constant corresponding to $\mathcal{O}$ (we can also consider perturbations in more than one coupling constant). At the same time thinking of the λ_a's as coordinates in the space of all Hamiltonians this corresponds to moving slightly away from the position of the fixed point.

The question to decide now is in which direction the renormalization group flow will take the perturbed system. Working at first order in $\delta\lambda$ there are three possibilities:

- The renormalization group flow takes the system back to the fixed point. In this case the corresponding interaction $\mathcal{O}$ is called *irrelevant*.
- $\mathcal{R}$ takes the system away from the fixed point. If this is what happens the interaction is called *relevant*.
- It is possible that the perturbation actually does not take the system away from the fixed point at first order in $\delta\lambda$. In this case the interaction is said to be *marginal* and it is necessary to go to higher orders in $\delta\lambda$ in order to decide whether the system moves to or away the fixed point, or whether we have a family of fixed points.

Therefore we can picture the action of the renormalization group transformation as a flow in the space of coupling constants. In Fig. 17 we have depicted an example of such a flow in the case of a system with two coupling constants λ_1 and λ_2. In this example we find two fixed points, one at the origin O and another at F for a finite value of the couplings. The arrows indicate the direction in which the renormalization group flow acts. The free theory at $\lambda_1 = \lambda_2 = 0$ is a stable fix point since any perturbation $\delta\lambda_1, \delta\lambda_2 > 0$ makes the theory flow back to the free theory at long distances. On the other hand, the fixed point F is stable with respect to certain type of perturbations (along the line with incoming arrows) whereas for any other perturbations the system flows either to the free theory at the origin or to a theory with infinite values for the couplings.

Quantum field theory. Let us see now how these ideas of the renormalization group apply to Field Theory. Let us begin with a quantum field theory defined by the Lagrangian

$$\mathcal{L}[\phi_a] = \mathcal{L}_0[\phi_a] + \sum_i g_i \mathcal{O}_i[\phi_a], \tag{404}$$

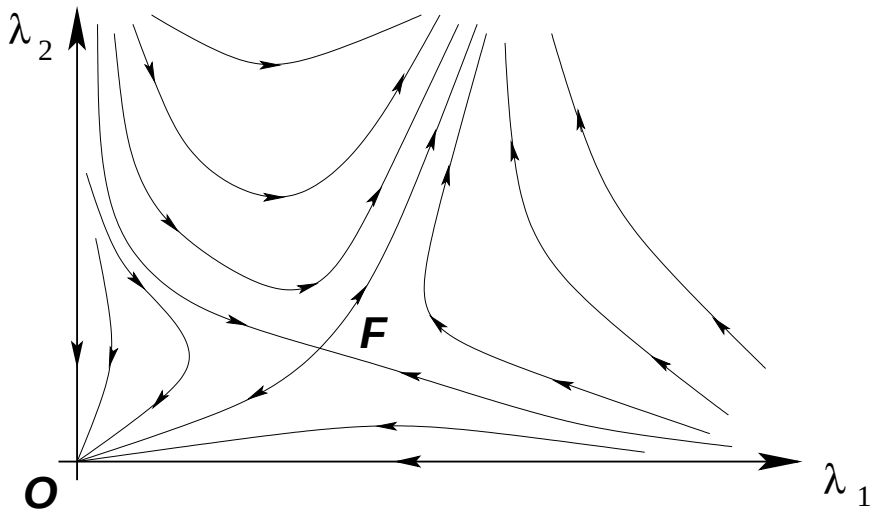


Fig. 17: Example of a renormalization group flow.

where $\mathcal{L}_0[\phi_a]$ is the kinetic part of the Lagrangian and g_i are the coupling constants associated with the operators $\mathcal{O}_i[\phi_a]$. In order to make sense of the quantum theory we introduce a cutoff in momenta Λ. In principle we include all operators $\mathcal{O}_i$ compatible with the symmetries of the theory.

In section 8.2 we saw how in the cases of QED and QCD, the value of the coupling constant changed with the scale from its value at the scale Λ. We can understand now this behavior along the lines of the analysis presented above for the Ising model. If we would like to compute the effective dynamics of the theory at an energy scale $\mu < \Lambda$ we only have to integrate out all physical models with energies between the cutoff Λ and the scale of interest μ. This is analogous to what we did in the Ising model by replacing the original spins by the block spins. In the case of field theory the effective action $S[\phi_a, \mu]$ at scale μ can be written in the language of functional integration as

$$e^{iS[\phi'_a,\mu]} = \int_{\mu<p<\Lambda} \prod_a \mathcal{D}\phi_a \, e^{iS[\phi_a,\Lambda]}. \tag{405}$$

Here $S[\phi_a, \Lambda]$ is the action at the cutoff scale

$$S[\phi_a, \Lambda] = \int d^4x \left\{ \mathcal{L}_0[\phi_a] + \sum_i g_i(\Lambda)\mathcal{O}_i[\phi_a] \right\} \tag{406}$$

and the functional integral in Eq. (405) is carried out only over the field modes with momenta in the range $\mu < p < \Lambda$. The action resulting from integrating out the physics at the intermediate scales between Λ and μ depends not on the original field variable ϕ_a but on some renormalized field ϕ'_a. At the same time the couplings $g_i(\mu)$ differ from their values at the cutoff scale $g_i(\Lambda)$. This is analogous to what we learned in the Ising model: by integrating out short distance physics we ended up with a new Hamiltonian depending on renormalized effective spin variables and with renormalized values for the coupling constants. Therefore the resulting effective action at scale μ can be written as

$$S[\phi'_a, \mu] = \int d^4x \left\{ \mathcal{L}_0[\phi'_a] + \sum_i g_i(\mu)\mathcal{O}_i[\phi'_a] \right\}. \tag{407}$$

This Wilsonian interpretation of renormalization sheds light to what in section 8.1 might have looked just a smart way to get rid of the infinities. The running of the coupling constant with the energy scale can be understood now as a way of incorporating into an effective action at scale μ the effects of field excitations at higher energies $E > \mu$.

As in statistical mechanics there are also quantum field theories that are fixed points of the renormalization group flow, i.e. whose coupling constants do not change with the scale. We have encountered them already in Section 8.2 when studying the properties of the beta function. The most trivial example of such theories are massless free quantum field theories, but there are also examples of four-dimensional interacting quantum field theories which are scale invariant. Again we can ask the question of what happens when a scale invariant theory is perturbed with some operator. In general the perturbed theory is not scale invariant anymore but we may wonder whether the perturbed theory flows at low energies towards or away the theory at the fixed point.

In quantum field theory this can be decided by looking at the canonical dimension $d[\mathcal{O}]$ of the operator $\mathcal{O}[\phi_a]$ used to perturb the theory at the fixed point. In four dimensions the three possibilities are defined by:

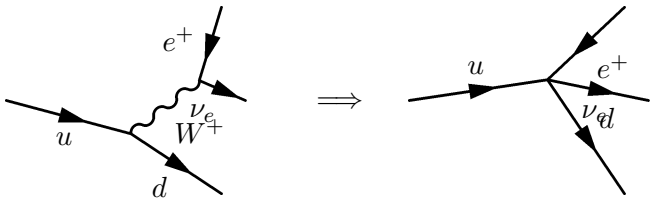
- $d[\mathcal{O}] > 4$: irrelevant perturbation. The running of the coupling constants takes the theory back to the fixed point.
- $d[\mathcal{O}] < 4$: relevant perturbation. At low energies the theory flows away from the scale-invariant theory.
- $d[\mathcal{O}] = 4$: marginal deformation. The direction of the flow cannot be decided only on dimensional grounds.

As an example, let us consider first a massless fermion theory perturbed by a four-fermion interaction term

$$\mathcal{L} = i\overline{\psi}\partial\!\!\!/\psi - \frac{1}{M^2}(\overline{\psi}\psi)^2. \tag{408}$$

This is indeed a perturbation by an irrelevant operator, since in four-dimensions $[\psi] = \frac{3}{2}$. Interactions generated by the extra term are suppressed at low energies since typically their effects are weighted by the dimensionless factor $\frac{E^2}{M^2}$, where E is the energy scale of the process. This means that as we try to capture the relevant physics at lower and lower energies the effect of the perturbation is weaker and weaker rendering in the infrared limit $E \to 0$ again a free theory. Hence, the irrelevant perturbation in (408) makes the theory flow back to the fixed point.

On the other hand relevant operators dominate the physics at low energies. This is the case, for example, of a mass term. As we lower the energy the mass becomes more important and once the energy goes below the mass of the field its dynamics is completely dominated by the mass term. This is, for example, how Fermi's theory of weak interactions emerges from the Standard Model at energies below the mass of the $W^{\pm}$ boson



At energies below $M_W = 80.4$ GeV the dynamics of the W^+ boson is dominated by its mass term and therefore becomes nonpropagating, giving rise to the effective four-fermion Fermi theory.

To summarize our discussion so far, we found that while relevant operators dominate the dynamics in the infrared, taking the theory away from the fixed point, irrelevant perturbations become suppressed in the same limit. Finally we consider the effect of marginal operators. As an example we take the interaction term in massless QED, $\mathcal{O} = \overline{\psi}\gamma^\mu\psi\, A_\mu$. Taking into account that in $d = 4$ the dimension of the electromagnetic potential is $[A_\mu] = 1$ the operator $\mathcal{O}$ is a marginal perturbation. In order to decide whether the fixed point theory

$$\mathcal{L}_0 = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\overline{\psi}\not{D}\,\psi \tag{409}$$

is restored at low energies or not we need to study the perturbed theory in more detail. This we have done in section 8.1 where we learned that the effective coupling in QED decreases at low energies. Then we conclude that the perturbed theory flows towards the fixed point in the infrared.

As an example of a marginal operator with the opposite behavior we can write the Lagrangian for a SU(N_c) gauge theory, $\mathcal{L} = -\frac{1}{4} F^a_{\mu\nu} F^{a\,\mu\nu}$, as

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4}\left(\partial_\mu A^a_\nu - \partial_\nu A^a_\mu\right)\left(\partial^\mu A^{a\,\nu} - \partial^\nu A^{a\,\mu}\right) - 4g f^{abc} A^a_\mu A^b_\nu\, \partial^\mu A^{c\,\nu} \\ &+ g^2 f^{abc} f^{ade} A^b_\mu A^c_\nu A^{d\,\mu} A^{e\,\nu} \equiv \mathcal{L}_0 + \mathcal{O}_g, \end{aligned} \tag{410}$$

i.e. a marginal perturbation of the free theory described by $\mathcal{L}_0$, which is obviously a fixed point under renormalization group transformations. Unlike the case of QED we know that the full theory is asymptotically free, so the coupling constant grows at low energies. This implies that the operator $\mathcal{O}_g$ becomes more and more important in the infrared and therefore the theory flows away the fixed point in this limit.

It is very important to notice here that in the Wilsonian view the cutoff is not necessarily regarded as just some artifact to remove infinities but actually has a physical origin. For example in the case of Fermi's theory of β-decay there is a natural cutoff $\Lambda = M_W$ at which the theory has to be replaced by the Standard Model. In the case of the Standard Model itself the cutoff can be taken at Planck scale $\Lambda \simeq 10^{19}$ GeV or the Grand Unification scale $\Lambda \simeq 10^{16}$ GeV, where new degrees of freedom are expected to become relevant. The cutoff serves the purpose of cloaking the range of energies at which new physics has to be taken into account.

Provided that in the Wilsonian approach the quantum theory is always defined with a physical cutoff, there is no fundamental difference between renormalizable and nonrenormalizable theories. Actually, a renormalizable field theory, like the Standard Model, can generate nonrenormalizable operators at low energies such as the effective four-fermion interaction of Fermi's theory. They are not sources of any trouble if we are interested in the physics at scales much below the cutoff, $E \ll \Lambda$, since their contribution to the amplitudes will be suppressed by powers of $\frac{E}{\Lambda}$.

9 Special topics

9.1 Creation of particles by classical fields

Particle creation by a classical source. In a free quantum field theory the total number of particles contained in a given state of the field is a conserved quantity. For example, in the case of the quantum scalar field studied in section 3 we have that the number operator commutes with the Hamiltonian

$$\widehat{n} \equiv \int \frac{d^3k}{(2\pi)^3} \frac{1}{2\omega_k} \alpha^\dagger(\vec{k})\alpha(\vec{k}), \qquad [\widehat{H}, \widehat{n}] = 0. \tag{411}$$

This means that any states with a well-defined number of particle excitations will preserve this number at all times. The situation, however, changes as soon as interactions are introduced, since in this case particles can be created and/or destroyed as a result of the dynamics.

Another case in which the number of particles might change is if the quantum theory is coupled to a classical source. The archetypical example of such a situation is the Schwinger effect, in which a classical strong electric field produces the creation of electron-positron pairs out of the vacuum. However, before plunging into this more involved situation we can illustrate the relevant physics involved in the creation of particles by classical sources with the help of the simplest example: a free scalar field theory coupled to a classical external source $J(x)$. The action for such a theory can be written as

$$S = \int d^4x \left[\frac{1}{2}\partial_\mu\phi(x)\partial^\mu\phi(x) - \frac{m^2}{2}\phi(x)^2 + J(x)\phi(x)\right], \tag{412}$$

where $J(x)$ is a real function of the coordinates. Its identification with a classical source is obvious once we calculate the equations of motion

$$\left(\nabla^2 + m^2\right)\phi(x) = J(x). \tag{413}$$

Our plan is to quantize this theory but, unlike the case analyzed in section 3, now the presence of the source $J(x)$ makes the situation a bit more involved. The general solution to the equations of motion can be written in terms of the retarded Green function for the Klein-Gordon equation as

$$\phi(x) = \phi_0(x) + i\int d^4x'\, G_R(x-x')J(x'), \tag{414}$$

where $\phi_0(x)$ is a general solution to the homogeneous equation and

$$\begin{aligned} G_R(t,\vec{x}) &= \int \frac{d^4k}{(2\pi)^4}\frac{i}{k^2 - m^2 + i\epsilon\,\mathrm{sign}(k^0)}e^{-ik\cdot x} \\ &= i\,\theta(t)\int \frac{d^3k}{(2\pi)^3}\frac{1}{2\omega_k}\left(e^{-i\omega_k t + \vec{k}\cdot\vec{x}} - e^{i\omega_k t - i\vec{p}\cdot\vec{x}}\right), \end{aligned} \tag{415}$$

with $\theta(x)$ the Heaviside step function. The integration contour to evaluate the integral over p^0 surrounds the poles at $p^0 = \pm\omega_k$ from above. Since $G_R(t,\vec{x}) = 0$ for $t < 0$, the function $\phi_0(x)$ corresponds to the solution of the field equation at $t \to -\infty$, before the interaction with the external source[21]

To make the argument simpler we assume that $J(x)$ is switched on at $t = 0$, and only last for a time τ, that is

$$J(t,\vec{x}) = 0 \qquad \text{if } t < 0 \text{ or } t > \tau. \tag{416}$$

We are interested in a solution of (413) for times after the external source has been switched off, $t > \tau$. In this case the expression (415) can be written in terms of the Fourier modes $\widetilde{J}(\omega,\vec{k})$ of the source as

$$\phi(t,\vec{x}) = \phi_0(x) + i\int \frac{d^3k}{(2\pi)^3}\frac{1}{2\omega_k}\left[\widetilde{J}(\omega_k,\vec{k})e^{-i\omega_k t + i\vec{k}\cdot\vec{x}} - \widetilde{J}(\omega_k,\vec{k})^* e^{i\omega_k t - i\vec{k}\cdot\vec{x}}\right]. \tag{417}$$

On the other hand, the general solution $\phi_0(x)$ has been already computed in Eq. (77). Combining this result with Eq. (417) we find the following expression for the late time general solution to the Klein-Gordon equation in the presence of the source

$$\begin{aligned} \phi(t,x) &= \int \frac{d^3k}{(2\pi)^3}\frac{1}{\sqrt{2\omega_k}}\left\{\left[\alpha(\vec{k}) + \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k})\right]e^{-i\omega_k t + i\vec{k}\cdot\vec{x}}\right. \\ &+ \left.\left[\alpha^*(\vec{k}) - \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k})^*\right]e^{i\omega_k t - i\vec{k}\cdot\vec{x}}\right\}. \end{aligned} \tag{418}$$

We should not forget that this is a solution valid for times $t > \tau$, i.e. once the external source has been disconnected. On the other hand, for $t < 0$ we find from Eqs. (414) and (415) that the general solution is given by Eq. (77).

Now we can proceed to quantize the theory. The conjugate momentum $\pi(x) = \partial_0\phi(x)$ can be computed from Eqs. (77) and (418). Imposing the canonical equal time commutation relations (74) we find that $\alpha(\vec{k})$, $\alpha^\dagger(\vec{k})$ satisfy the creation-annihilation algebra (51). From our previous calculation we find that for $t > \tau$ the expansion of the operator $\phi(x)$ in terms of the creation-annihilation operators $\alpha(\vec{k})$, $\alpha^\dagger(\vec{k})$ can be obtained from the one for $t < 0$ by the replacement

$$\alpha(\vec{k}) \quad\longrightarrow\quad \beta(\vec{k}) \equiv \alpha(\vec{k}) + \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k}),$$

[21] We could have taken instead the advanced propagator $G_A(x)$ in which case $\phi_0(x)$ would correspond to the solution to the equation at large times, after the interaction with $J(x)$.

$$\alpha^\dagger(\vec{k}) \quad \longrightarrow \quad \beta^\dagger(\vec{k}) \equiv \alpha^\dagger(\vec{k}) - \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k})^*. \tag{419}$$

Actually, since $\widetilde{J}(\omega_k,\vec{k})$ is a c-number, the operators $\beta(\vec{k})$, $\beta^\dagger(\vec{k})$ satisfy the same algebra as $\alpha(\vec{k})$, $\alpha^\dagger(\vec{k})$ and therefore can be interpreted as well as a set of creation-annihilation operators. This means that we can define two vacuum states, $|0_-\rangle$, $|0_+\rangle$ associated with both sets of operators

$$\left.\begin{array}{l}\alpha(\vec{k})|0_-\rangle = 0 \\ \\ \beta(\vec{k})|0_+\rangle = 0\end{array}\right\} \qquad \forall\ \vec{k}. \tag{420}$$

For an observer at $t < 0$, $\alpha(\vec{k})$ and $\alpha(\vec{k})$ are the natural set of creation-annihilation operators in terms of which to expand the field operator $\phi(x)$. After the usual zero-point energy subtraction the Hamiltonian is given by

$$\widehat{H}^{(-)} = \frac{1}{2}\int \frac{d^3k}{(2\pi)^3}\,\alpha^\dagger(\vec{k})\alpha(\vec{k}) \tag{421}$$

and the ground state of the spectrum for this observer is the vacuum $|0_-\rangle$. At the same time, a second observer at $t > \tau$ will also see a free scalar quantum field (the source has been switched off at $t = \tau$) and consequently will expand ϕ in terms of the second set of creation-annihilation operators $\beta(\vec{k})$, $\beta^\dagger(\vec{k})$. In terms of this operators the Hamiltonian is written as

$$\widehat{H}^{(+)} = \frac{1}{2}\int \frac{d^3k}{(2\pi)^3}\,\beta^\dagger(\vec{k})\beta(\vec{k}). \tag{422}$$

Then for this late-time observer the ground state of the Hamiltonian is the second vacuum state $|0_+\rangle$.

In our analysis we have been working in the Heisenberg picture, where states are time-independent and the time dependence comes in the operators. Therefore the states of the theory are globally defined. Suppose now that the system is in the "in" ground state $|0_-\rangle$. An observer at $t < 0$ will find that there are no particles

$$\widehat{n}^{(-)}|0_-\rangle = 0. \tag{423}$$

However the late-time observer will find that the state $|0_-\rangle$ contains an average number of particles given by

$$\langle 0_-|\widehat{n}^{(+)}|0_-\rangle = \int \frac{d^3k}{(2\pi)^3}\frac{1}{2\omega_k}\left|\widetilde{J}(\omega_k,\vec{k})\right|^2. \tag{424}$$

Moreover, $|0_-\rangle$ is no longer the ground state for the "out" observer. On the contrary, this state have a vacuum expectation value for $\widehat{H}^{(+)}$

$$\langle 0_-|\widehat{H}^{(+)}|0_-\rangle = \frac{1}{2}\int \frac{d^3k}{(2\pi)^3}\left|\tilde{J}(\omega_k,\vec{k})\right|^2. \tag{425}$$

The key to understand what is going on here lies in the fact that the external source breaks the invariance of the theory under space-time translations. In the particular case we have studied here where $J(x)$ has support over a finite time interval $0 < t < \tau$, this implies that the vacuum is not invariant under time translations, so observers at different times will make different choices of vacuum that will not necessarily agree with each other. This is clear in our example. An observer in $t < \tau$ will choose the vacuum to be the lowest energy state of her Hamiltonian, $|0_-\rangle$. On the other hand, the second observer at late times $t > \tau$ will naturally choose $|0_+\rangle$ as the vacuum. However, for this second observer, the

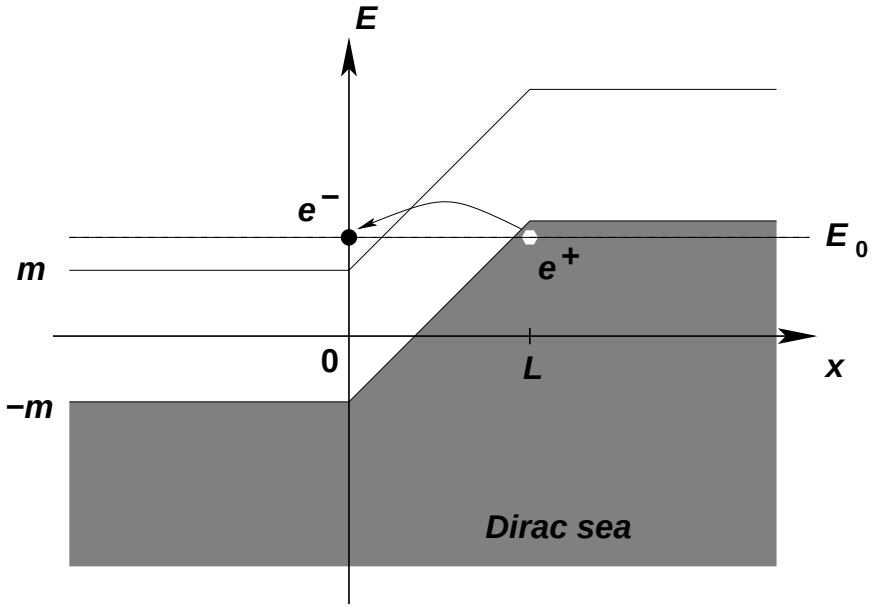


Fig. 18: Pair creation by a electric field in the Dirac sea picture.

state $|0_-\rangle$ is not the vacuum of his Hamiltonian, but actually an excited state that is a superposition of states with well-defined number of particles. In this sense it can be said that the external source has the effect of creating particles out of the "in" vacuum. Besides, this breaking of time translation invariance produces a violation in the energy conservation as we see from Eq. (425). Particles are actually created from the energy pumped into the system by the external source.

The Schwinger effect. A classical example of creation of particles by a external field was pointed out by Schwinger [44] and consists of the creation of electron-positron pairs by a strong electric field. In order to illustrate this effect we are going to follow a heuristic argument based on the Dirac sea picture and the WKB approximation.

In the absence of an electric field the vacuum state of a spin-$\frac{1}{2}$ field is constructed by filling all the negative energy states as depicted in Fig. 2. Let us now connect a constant electric field $\vec{\mathcal{E}} = \mathcal{E}\vec{u}_x$ in the range $0 < x < L$ created by a electrostatic potential

$$V(\vec{r}) = \begin{cases} 0 & x < 0 \\ -\mathcal{E}x & 0 < x < L \\ -\mathcal{E}L & x > L \end{cases} \tag{426}$$

After the field has been switched on, the Dirac sea looks like in Fig. 18. In particular we find that if $e\mathcal{E}L > 2m$ there are negative energy states at $x > L$ with the same energy as the positive energy states in the region $x < 0$. Therefore it is possible for an electron filling a negative energy state with energy close to $-2m$ to tunnel through the forbidden region into a positive energy state. The interpretation of such a process is the production of an electron-positron pair out of the electric field.

We can compute the rate at which such pairs are produced by using the WKB approximation. Focusing for simplicity on an electron on top of the Fermi surface near $x = L$ with energy E_0, the transmission coefficient in this approximation is given by[22]

$$T_{\rm WKB} \;=\; \exp\left[-2\int_{\frac{1}{e\mathcal{E}}\left(E_0-\sqrt{m^2+\vec{p}_T^{\,2}}\right)}^{\frac{1}{e\mathcal{E}}\left(E_0+\sqrt{m^2+\vec{p}_T^{\,2}}\right)} dx\,\sqrt{m^2-\left[E_0-e\mathcal{E}(x-x_0)\right]^2+\vec{p}_T^{\,2}}\right]$$

[22]Notice that the electron satisfy the relativistic dispersion relation $E = \sqrt{\vec{p}^{\,2}+m^2}+V$ and therefore $-p_x^2 = m^2-(E-V)^2+\vec{p}_T^{\,2}$. The integration limits are set by those values of x at which $p_x = 0$.

$$= \quad \exp\left[-\frac{\pi}{e\mathcal{E}}\left(\vec{p}_T^{\,2}+m^2\right)\right], \tag{427}$$

where $p_T^2 \equiv p_y^2 + p_z^2$. This gives the transition probability per unit time and per unit cross section $dydz$ for an electron in the Dirac sea with transverse momentum $\vec{p}_T$ and energy E_0. To get the total probability per unit time and per unit volume we have to integrate over all possible values of $\vec{p}_T$ and E_0. Actually, in the case of the energy, because of the relation between E_0 and the coordinate x at which the particle penetrates into the barrier we can write $\frac{dE_0}{2\pi} = \frac{e\mathcal{E}}{2\pi}dx$ and the total probability per unit time and per unit volume for the creation of a pair is given by

$$W = 2\left(\frac{e\mathcal{E}}{2\pi}\right)\int \frac{d^2p_T}{(2\pi)^2} e^{-\frac{\pi}{e\mathcal{E}}\left(\vec{p}_T^{\,2}+m^2\right)} = \frac{e^2\mathcal{E}^2}{4\pi^3} e^{-\frac{\pi\, m^2}{e\mathcal{E}}}, \tag{428}$$

where the factor of 2 accounts for the two polarizations of the electron.

Then production of electron-positron pairs is exponentially suppressed and it is only sizeable for strong electric fields. To estimate its order of magnitude it is useful to restore the powers of c and $\hbar$ in (428)

$$W = \frac{e^2\mathcal{E}^2}{4\pi^3 c\hbar^2} e^{-\frac{\pi\, m^2 c^3}{\hbar e\mathcal{E}}} \tag{429}$$

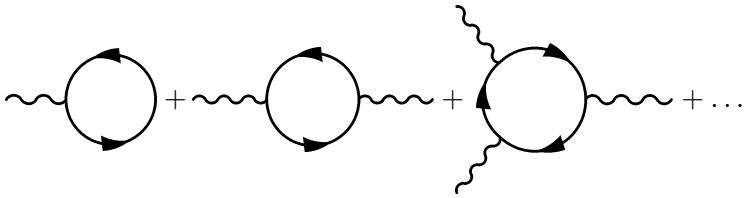
The exponential suppression of the pair production disappears when the electric field reaches the critical value $\mathcal{E}_{\rm crit}$ at which the exponent is of order one

$$\mathcal{E}_{\rm crit} = \frac{m^2c^3}{\hbar e} \simeq 1.3\times 10^{16}\,{\rm V\,cm^{-1}}. \tag{430}$$

This is indeed a very strong field which is extremely difficult to produce. A similar effect, however, takes place also in a time-varying electric field [45] and there is the hope that pair production could be observed in the presence of the alternating electric field produced by a laser.

The heuristic derivation that we followed here can be made more precise in QED. There the decay of the vacuum into electron-positron pairs can be computed from the imaginary part of the effective action $\Gamma[A_\mu]$ in the presence of a classical gauge potential A_μ

$i\Gamma[A_\mu] \quad \equiv$



$$= \quad \log\ \det\left[1 - ie\not{A}\frac{1}{i\not{\partial} - m}\right]. \tag{431}$$

This determinant can be computed using the standard heat kernel techniques. The probability of pair production is proportional to the imaginary part of $i\Gamma[A_\mu]$ and gives

$$W = \frac{e^2\mathcal{E}^2}{4\pi^3}\sum_{n=1}^{\infty}\frac{1}{n^2}e^{-n\frac{\pi\, m^2}{e\mathcal{E}}}. \tag{432}$$

Our simple argument based on tunneling in the Dirac sea gave only the leading term of Schwinger's result (432). The remaining terms can be also captured in the WKB approximation by taking into account the probability of production of several pairs, i.e. the tunneling of more than one electron through the barrier.

Here we have illustrated the creation of particles by semiclassical sources in Quantum Field Theory using simple examples. Nevertheless, what we learned has important applications to the study of

quantum fields in curved backgrounds. In Quantum Field Theory in Minkowski space-time the vacuum state is invariant under the Poincaré group and this, together with the covariance of the theory under Lorentz transformations, implies that all inertial observers agree on the number of particles contained in a quantum state. The breaking of such invariance, as happened in the case of coupling to a time-varying source analyzed above, implies that it is not possible anymore to define a state which would be recognized as the vacuum by all observers.

This is precisely the situation when fields are quantized on curved backgrounds. In particular, if the background is time-dependent (as it happens in a cosmological setup or for a collapsing star) different observers will identify different vacuum states. As a consequence what one observer call the vacuum will be full of particles for a different observer. This is precisely what is behind the phenomenon of Hawking radiation [46]. The emission of particles by a physical black hole formed from gravitational collapse of a star is the consequence of the fact that the vacuum state in the asymptotic past contain particles for an observer in the asymptotic future. As a consequence, a detector located far away from the black hole detects a stream of thermal radiation with temperature

$$T_{\rm Hawking} = \frac{\hbar c^3}{8\pi G_N\, k\, M} \tag{433}$$

where M is the mass of the black hole, G_N is Newton's constant and k is Boltzmann's constant. There are several ways in which this results can be obtained. A more heuristic way is perhaps to think of this particle creation as resulting from quantum tunneling of particles across the potential barrier posed by gravity [47].

9.2 Supersymmetry

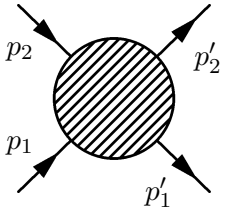
One of the things that we have learned in our journey around the landscape of Quantum Field Theory is that our knowledge of the fundamental interactions in Nature is based on the idea of symmetry, and in particular gauge symmetry. The Lagrangian of the Standard Model can be written just including all possible renormalizable terms (i.e. with canonical dimension smaller o equal to 4) compatible with the gauge symmetry SU(3)×SU(2)×U(1) and Poincaré invariance. All attempts to go beyond start with the question of how to extend the symmetries of the Standard Model.

As explained in Section 5.1, in a quantum field theoretical description of the interaction of elementary particles the basic observable quantity to compute is the scattering or S-matrix giving the probability amplitude for the scattering of a number of incoming particles with a certain momentum into some final products

$$\mathcal{A}(\text{in} \longrightarrow \text{out}) = \langle \vec{p}_1{}', \ldots; \text{out} | \vec{p}_1, \ldots; \text{in} \rangle. \tag{434}$$

An explicit symmetry of the theory has to be necessarily a symmetry of the S-matrix. Hence it is fair to ask what is the largest symmetry of the S-matrix.

Let us ask this question in the simple case of the scattering of two particles with four-momenta p_1 and p_2 in the t-channel



We will make the usual assumptions regarding positivity of the energy and analyticity. Invariance of the theory under the Poincaré group implies that the amplitude can only depend on the scattering angle ϑ through

$$t = (p_1' - p_1)^2 = 2\left(m_1^2 - p_1 \cdot p_1'\right) = 2\left(m_1^2 - E_1 E_1' + |\vec{p}_1||\vec{p}_1{}'|\cos\vartheta\right). \tag{435}$$

If there would be any extra bosonic symmetry of the theory it would restrict the scattering angle to a set of discrete values. In this case the S-matrix cannot be analytic since it would vanish everywhere except for the discrete values selected by the extra symmetry.

Actually, the only way to extend the symmetry of the theory without renouncing to the analyticity of the scattering amplitudes is to introduce "fermionic" symmetries, i.e. symmetries whose generators are anticommuting objects [48]. This means that in addition to the generators of the Poincaré group[23] P^μ, $M^{\mu\nu}$ and the ones for the internal gauge symmetries G, we can introduce a number of fermionic generators Q^I_a, $\overline{Q}_{\dot{a}\,I}$ ($I = 1, \ldots, \mathcal{N}$), where $\overline{Q}_{\dot{a}\,I} = (Q^I_a)^\dagger$. The most general algebra that these generators satisfy is the $\mathcal{N}$-extended supersymmetry algebra [49]

$$\begin{aligned} \{Q^I_a, \overline{Q}_{\dot{b}\,J}\} &= 2\sigma^\mu_{a\dot{b}} P_\mu \delta^I{}_J, \\ \{Q^I_a, Q^J_b\} &= 2\varepsilon_{ab}\mathcal{Z}^{IJ}, \qquad (436) \\ \{\overline{Q}^I_{\dot{a}}, \overline{Q}^J_{\dot{b}}\} &= 2\varepsilon_{\dot{a}\dot{b}}\overline{\mathcal{Z}}^{IJ}, \qquad (437) \end{aligned}$$

where $\mathcal{Z}^{IJ} \in \mathbb{C}$ commute with any other generator and satisfies $\mathcal{Z}^{IJ} = -\mathcal{Z}^{JI}$. Besides we have the commutators that determine the Poincaré transformations of the fermionic generators Q^I_a, $Q_{\dot{a}\,J}$

$$\begin{aligned} [Q^I_a, P^\mu] &= [\overline{Q}_{\dot{a}\,I}, P^\mu] = 0, \\ [Q^I_a, M^{\mu\nu}] &= \frac{1}{2}(\sigma^{\mu\nu})_a{}^b Q^I_b, \qquad (438) \\ [\overline{Q}_{a\,I}, M^{\mu\nu}] &= -\frac{1}{2}(\overline{\sigma}^{\mu\nu})_{\dot{a}}{}^{\dot{b}}\,\overline{Q}_{\dot{b}\,I}, \end{aligned}$$

where $\sigma^{0i} = -i\sigma^i$, $\sigma^{ij} = \varepsilon^{ijk}\sigma^k$ and $\overline{\sigma}^{\mu\nu} = (\sigma^{\mu\nu})^\dagger$. These identities simply mean that Q^I_a, $\overline{Q}_{\dot{a}\,J}$ transform respectively in the $(\frac{\mathbf{1}}{\mathbf{2}}, \mathbf{0})$ and $(\mathbf{0}, \frac{\mathbf{1}}{\mathbf{2}})$ representations of the Lorentz group.

We know that the presence of a global symmetry in a theory implies that the spectrum can be classified in multiplets with respect to that symmetry. In the case of supersymmetry start with the case case $\mathcal{N} = 1$ in which there is a single pair of supercharges Q_a, $\overline{Q}_{\dot{a}}$ satisfying the algebra

$$\{Q_a, \overline{Q}_{\dot{b}}\} = 2\sigma^\mu_{a\dot{b}} P_\mu, \qquad \{Q_a, Q_b\} = \{\overline{Q}_{\dot{a}}, \overline{Q}_{\dot{b}}\} = 0. \qquad (439)$$

Notice that in the $\mathcal{N} = 1$ case there is no possibility of having central charges.

We study now the representations of the supersymmetry algebra (439), starting with the massless case. Given a state $|k\rangle$ satisfying $k^2 = 0$, we can always find a reference frame where the four-vector k^μ takes the form $k^\mu = (E, 0, 0, E)$. Since the theory is Lorentz covariant we can obtain the representation of the supersymmetry algebra in this frame where the expressions are simpler. In particular, the right-hand side of the first anticommutator in Eq. (439) is given by

$$2\sigma^\mu_{a\dot{b}} P_\mu = 2(P^0 - \sigma^3 P^3) = \begin{pmatrix} 0 & 0 \\ 0 & 4E \end{pmatrix}. \qquad (440)$$

Therefore the algebra of supercharges in the massless case reduces to

$$\begin{aligned} \{Q_1, Q_1^\dagger\} &= \{Q_1, Q_2^\dagger\} = 0, \\ \{Q_2, Q_2^\dagger\} &= 4E. \qquad (441) \end{aligned}$$

The commutator $\{Q_1, Q_1^\dagger\} = 0$ implies that the action of Q_1 on any state gives a zero-norm state of the Hilbert space $\|Q_1|\Psi\rangle\| = 0$. If we want the theory to preserve unitarity we must eliminate these null

[23] The generators $M^{\mu\nu}$ are related with the ones for boost and rotations introduced in section 4.1 by $J^i \equiv M^{0i}$, $M^i = \frac{1}{2}\varepsilon^{ijk}M^{jk}$. In this section we also use the "dotted spinor" notation, in which spinors in the $(\frac{\mathbf{1}}{\mathbf{2}}, \mathbf{0})$ and $(\mathbf{0}, \frac{\mathbf{1}}{\mathbf{2}})$ representations of the Lorentz group are indicated respectively by undotted $(a, b, \ldots)$ and dotted $(\dot{a}, \dot{b}, \ldots)$ indices.

states from the spectrum. This is equivalent to setting $Q_1 \equiv 0$. On the other hand, in terms of the second generator Q_2 we can define the operators

$$a = \frac{1}{2\sqrt{E}} Q_2, \qquad a^\dagger = \frac{1}{2\sqrt{E}} Q_2^\dagger, \tag{442}$$

which satisfy the algebra of a pair of fermionic creation-annihilation operators, $\{a, a^\dagger\} = 1$, $a^2 = (a^\dagger)^2 = 0$. Starting with a vacuum state $a|\lambda\rangle = 0$ with helicity λ we can build the massless multiplet

$$|\lambda\rangle, \qquad |\lambda + \tfrac{1}{2}\rangle \equiv a^\dagger |\lambda\rangle. \tag{443}$$

Here we consider two important cases:

– Scalar multiplet: we take the vacuum state to have zero helicity $|0^+\rangle$ so the multiplet consists of a scalar and a helicity-$\frac{1}{2}$ state

$$|0^+\rangle, \qquad |\tfrac{1}{2}\rangle \equiv a^\dagger |0^+\rangle. \tag{444}$$

However, this multiplet is not invariant under the CPT transformation which reverses the sign of the helicity of the states. In order to have a CPT-invariant theory we have to add to this multiplet its CPT-conjugate which can be obtain from a vacuum state with helicity $\lambda = -\frac{1}{2}$

$$|0^-\rangle, \qquad |-\tfrac{1}{2}\rangle. \tag{445}$$

Putting them together we can combine the two zero helicity states with the two fermionic ones into the degrees of freedom of a complex scalar field and a Weyl (or Majorana) spinor.

– Vector multiplet: now we take the vacuum state to have helicity $\lambda = \frac{1}{2}$, so the multiplet contains also a massless state with helicity $\lambda = 1$

$$|\tfrac{1}{2}\rangle, \qquad |1\rangle \equiv a^\dagger |\tfrac{1}{2}\rangle. \tag{446}$$

As with the scalar multiplet we add the CPT conjugated obtained from a vacuum state with helicity $\lambda = -1$

$$|-\tfrac{1}{2}\rangle, \qquad |-1\rangle, \tag{447}$$

which together with (446) give the propagating states of a gauge field and a spin-$\frac{1}{2}$ gaugino.

In both cases we see the trademark of supersymmetric theories: the number of bosonic and fermionic states within a multiplet are the same.

In the case of extended supersymmetry we have to repeat the previous analysis for each supersymmetry charge. At the end, we have $\mathcal{N}$ sets of fermionic creation-annihilation operators $\{a^I, a_I^\dagger\} = \delta^I{}_J$, $(a_I)^2 = (a_I^\dagger)^2 = 0$. Let us work out the case of $\mathcal{N} = 8$ supersymmetry. Since for several reasons we do not want to have states with helicity larger than 2, we start with a vacuum state $|-2\rangle$ of helicity $\lambda = -2$. The rest of the states of the supermultiplet are obtained by applying the eight different creation operators $a_I^\dagger$ to the vacuum:

$$\begin{aligned}
\lambda = 2: &\qquad a_1^\dagger \dots a_8^\dagger |-2\rangle &&\qquad \binom{8}{8} = 1 \text{ state},\\
\lambda = \frac{3}{2}: &\qquad a_{I_1}^\dagger \dots a_{I_7}^\dagger |-2\rangle &&\qquad \binom{8}{7} = 8 \text{ states},\\
\lambda = 1: &\qquad a_{I_1}^\dagger \dots a_{I_6}^\dagger |-2\rangle &&\qquad \binom{8}{6} = 28 \text{ states},
\end{aligned}$$

$$\begin{aligned}
\lambda = \frac{1}{2}: &\quad a^\dagger_{I_1} \dots a^\dagger_{I_5} |-2\rangle && \binom{8}{5} = 56 \text{ states}, \\
\lambda = 0: &\quad a^\dagger_{I_1} \dots a^\dagger_{I_4} |-2\rangle && \binom{8}{4} = 70 \text{ states}, \\
\lambda = -\frac{1}{2}: &\quad a^\dagger_{I_1} a^\dagger_{I_2} a^\dagger_{I_3} |-2\rangle && \binom{8}{3} = 56 \text{ states}, \\
\lambda = -1: &\quad a^\dagger_{I_1} a^\dagger_{I_2} |-2\rangle && \binom{8}{2} = 28 \text{ states}, \\
\lambda = -\frac{3}{2}: &\quad a^\dagger_{I_1} |-2\rangle && \binom{8}{1} = 8 \text{ states}, \\
\lambda = -2: &\quad |-2\rangle && 1 \text{ state}.
\end{aligned} \tag{448}$$

Putting together the states with opposite helicity we find that the theory contains:

- 1 spin-2 field $g_{\mu\nu}$ (a graviton),
- 8 spin-$\frac{3}{2}$ gravitino fields ψ^I_μ,
- 28 gauge fields $A^{[IJ]}_\mu$,
- 56 spin-$\frac{1}{2}$ fermions $\psi^{[IJK]}$,
- 70 scalars $\phi^{[IJKL]}$,

where by $[IJ...]$ we have denoted that the indices are antisymmetrized. We see that, unlike the massless multiplets of $\mathcal{N} = 1$ supersymmetry studied above, this multiplet is CPT invariant by itself. As in the case of the massless $\mathcal{N} = 1$ multiplet, here we also find as many bosonic as fermionic states:

$$\begin{aligned}
\text{bosons:} &\quad 1 + 28 + 70 + 28 + 1 = 128 \quad \text{states}, \\
\text{fermions:} &\quad 8 + 56 + 56 + 8 = 128 \quad \text{states}.
\end{aligned}$$

Now we study briefly the case of massive representations $|k\rangle$, $k^2 = M^2$. Things become simpler if we work in the rest frame where $P^0 = M$ and the spatial components of the momentum vanish. Then, the supersymmetry algebra becomes:

$$\{Q^I_a, \overline{Q}_{\dot{b}\,J}\} = 2M\delta_{a\dot{b}}\delta^I{}_J. \tag{449}$$

We proceed now in a similar way to the massless case by defining the operators

$$a^I_a \equiv \frac{1}{\sqrt{2M}} Q^I_a, \qquad a^\dagger_{\dot{a}\,I} \equiv \frac{1}{\sqrt{2M}} \overline{Q}_{\dot{a}\,I}. \tag{450}$$

The multiplets are found by choosing a vacuum state with a definite spin. For example, for $\mathcal{N} = 1$ and taking a spin-0 vacuum $|0\rangle$ we find three states in the multiplet transforming irreducibly with respect to the Lorentz group:

$$|0\rangle, \qquad a^\dagger_{\dot{a}}|0\rangle, \qquad \varepsilon^{\dot{a}\dot{b}} a^\dagger_{\dot{a}} a^\dagger_{\dot{b}} |0\rangle, \tag{451}$$

which, once transformed back from the rest frame, correspond to the physical states of two spin-0 bosons and one spin-$\frac{1}{2}$ fermion. For $\mathcal{N}$-extended supersymmetry the corresponding multiplets can be worked out in a similar way.

The equality between bosonic and fermionic degrees of freedom is at the root of many of the interesting properties of supersymmetric theories. For example, in section 4 we computed the divergent vacuum energy contributions for each real bosonic or fermionic propagating degree of freedom is[24]

$$E_{\text{vac}} = \pm \frac{1}{2} \delta(\vec{0}) \int d^3p\, \omega_p, \tag{452}$$

[24]For a boson, this can be read off Eq. (80). In the case of fermions, the result of Eq. (134) gives the vacuum energy contribution of the four real propagating degrees of freedom of a Dirac spinor.

where the sign $\pm$ corresponds respectively to bosons and fermions. Hence, for a supersymmetric theory the vacuum energy contribution exactly cancels between bosons and fermions. This boson-fermion degeneracy is also responsible for supersymmetric quantum field theories being less divergent than non-supersymmetric ones.

Appendix: A crash course in Group Theory

In this Appendix we summarize some basic facts about Group Theory. Given a group G a representation of G is a correspondence between the elements of G and the set of linear operators acting on a vector space V, such that for each element of the group $g \in G$ there is a linear operator $D(g)$

$$D(g) : V \longrightarrow V \tag{453}$$

satisfying the group operations

$$D(g_1)D(g_2) = D(g_1 g_2), \qquad D(g_1^{-1}) = D(g_1)^{-1}, \qquad g_1, g_2 \in \mathcal{G}. \tag{454}$$

The representation $D(g)$ is irreducible if and only if the only operators $A : V \to V$ commuting with all the elements of the representation $D(g)$ are the ones proportional to the identity

$$[D(g), A] = 0, \ \ \forall g \qquad \Longleftrightarrow \qquad A = \lambda \mathbf{1}, \quad \lambda \in \mathbb{C} \tag{455}$$

More intuitively, we can say that a representation is irreducible if there is no proper subspace $U \subset V$ (i.e. $U \neq V$ and $U \neq \emptyset$) such that $D(g)U \subset U$ for every element $g \in G$.

Here we are specially interested in Lie groups whose elements are labelled by a number of continuous parameters. In mathematical terms this means that a Lie group is a manifold $\mathcal{M}$ together with an operation $\mathcal{M} \times \mathcal{M} \longrightarrow \mathcal{M}$ that we will call multiplication that satisfies the associativity property $g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3$ together with the existence of unity $g\mathbf{1} = \mathbf{1}g = g$,for every $g \in \mathcal{M}$ and inverse $gg^{-1} = g^{-1}g = \mathbf{1}$.

The simplest example of a Lie group is SO(2), the group of rotations in the plane. Each element $R(\theta)$ is labelled by the rotation angle θ, with the multiplication acting as $R(\theta_1)R(\theta_2) = R(\theta_1 + \theta_2)$. Because the angle θ is defined only modulo 2π, the manifold of SO(2) is a circumference S^1.

One of the interesting properties of Lie groups is that in a neighborhood of the identity element they can be expressed in terms of a set of generators T^a ($a = 1, \ldots, \dim G$) as

$$D(g) = \exp(-i\alpha_a T^a) \equiv \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \alpha_{a_1} \ldots \alpha_{a_n} T^{a_1} \ldots T^{a_n}, \tag{456}$$

where $\alpha_a \in \mathbb{C}$ are a set of coordinates of $\mathcal{M}$ in a neighborhood of $\mathbf{1}$. Because of the general Baker-Campbell-Haussdorf formula, the multiplication of two group elements is encoded in the value of the commutator of two generators, that in general has the form

$$[T^a, T^b] = i f^{abc} T^c, \tag{457}$$

where $f^{abc} \in \mathbb{C}$ are called the structure constants. The set of generators with the commutator operation form the Lie algebra associated with the Lie group. Hence, given a representation of the Lie algebra of generators we can construct a representation of the group by exponentiation (at least locally near the identity).

We illustrate these concept with some particular examples. For SU(2) each group element is labelled by three real number α_i, $i = 1, 2, 3$. We have two basic representations: one is the fundamental representation (or spin $\frac{1}{2}$) defined by

$$D_{\frac{1}{2}}(\alpha_i) = e^{-\frac{i}{2}\alpha_i \sigma^i}, \tag{458}$$

with σ^i the Pauli matrices. The second one is the adjoint (or spin 1) representation which can be written as

$$D_1(\alpha_i) = e^{-i\alpha_i J^i}, \tag{459}$$

where

$$J^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \qquad J^2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad J^3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{460}$$

Actually, J^i $(i = 1, 2, 3)$ generate rotations around the x, y and z axis respectively. Representations of spin $j \in \mathbb{N} + \frac{1}{2}$ can be also constructed with dimension

$$\dim D_j(g) = 2j + 1. \tag{461}$$

As a second example we consider SU(3). This group has two basic three-dimensional representations denoted by $\mathbf{3}$ and $\overline{\mathbf{3}}$ which in QCD are associated with the transformation of quarks and antiquarks under the color gauge symmetry SU(3). The elements of these representations can be written as

$$D_{\mathbf{3}}(\alpha^a) = e^{\frac{i}{2}\alpha^a\lambda_a}, \qquad D_{\overline{\mathbf{3}}}(\alpha^a) = e^{-\frac{i}{2}\alpha^a\lambda_a^T} \qquad (a = 1, \ldots, 8), \tag{462}$$

where λ_a are the eight hermitian Gell-Mann matrices

$$\begin{aligned}
\lambda_1 &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
\lambda_4 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \qquad \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \\
\lambda_7 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \qquad \lambda_8 = \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & -\frac{2}{\sqrt{3}} \end{pmatrix}.
\end{aligned} \tag{463}$$

Hence the generators of the representations $\mathbf{3}$ and $\overline{\mathbf{3}}$ are given by

$$T^a(\mathbf{3}) = \frac{1}{2}\lambda_a, \qquad T^a(\overline{\mathbf{3}}) = -\frac{1}{2}\lambda_a^T. \tag{464}$$

Irreducible representations can be classified in three groups: real, complex and pseudoreal.

– Real representations: a representation is said to be real if there is a *symmetric matrix* S which acts as intertwiner between the generators and their complex conjugates

$$\overline{T}^a = -ST^aS^{-1}, \qquad S^T = S. \tag{465}$$

This is for example the case of the adjoint representation of SU(2) generated by the matrices (460)

– Pseudoreal representations: are the ones for which an *antisymmetric matrix* S exists with the property

$$\overline{T}^a = -ST^aS^{-1}, \qquad S^T = -S. \tag{466}$$

As an example we can mention the spin-$\frac{1}{2}$ representation of SU(2) generated by $\frac{1}{2}\sigma^i$.

– Complex representations: finally, a representation is complex if the generators and their complex conjugate are not related by a similarity transformation. This is for instance the case of the two three-dimensional representations $\mathbf{3}$ and $\overline{\mathbf{3}}$ of SU(3).

There are a number of invariants that can be constructed associated with an irreducible representation R of a Lie group G and that can be used to label such a representation. If T_R^a are the generators in a certain representation R of the Lie algebra, it is easy to see that the matrix $\sum_{a=1}^{\dim G} T_R^a T_R^a$ commutes with every generator T_R^a. Therefore, because of Schur's lemma, it has to be proportional to the identity[25]. This defines the Casimir invariant $C_2(R)$ as

$$\sum_{a=1}^{\dim G} T_R^a T_R^a = C_2(R)\mathbf{1}. \tag{467}$$

A second invariant $T_2(R)$ associated with a representation R can also be defined by the identity

$$\mathrm{Tr}\, T_R^a T_R^b = T_2(R)\delta^{ab}. \tag{468}$$

Actually, taking the trace in Eq. (467) and combining the result with (468) we find that both invariants are related by the identity

$$C_2(R)\dim R = T_2(R)\dim G, \tag{469}$$

with $\dim R$ the dimension of the representation R.

These two invariants appear frequently in quantum field theory calculations with nonabelian gauge fields. For example $T_2(R)$ comes about as the coefficient of the one-loop calculation of the beta-function for a Yang-Mills theory with gauge group G. In the case of SU(N), for the fundamental representation, we find the values

$$C_2(\text{fund}) = \frac{N^2-1}{2N}, \qquad T_2(\text{fund}) = \frac{1}{2}, \tag{470}$$

whereas for the adjoint representation the results are

$$C_2(\text{adj}) = N, \qquad T_2(\text{adj}) = N. \tag{471}$$

A third invariant $A(R)$ is specially important in the calculation of anomalies. As discussed in section (7), the chiral anomaly in gauge theories is proportional to the group-theoretical factor $\mathrm{Tr}\left[T_R^a\{T_R^b, T_R^c\}\right]$ This leads us to define $A(R)$ as

$$\mathrm{Tr}\,\left[T_R^a\{T_R^b, T_R^c\}\right] = A(R)d^{abc}, \tag{472}$$

where d^{abc} is symmetric in its three indices and does not depend on the representation. Therefore, the cancellation of anomalies in a gauge theory with fermions transformed in the representation R of the gauge group is guaranteed if the corresponding invariant $A(R)$ vanishes.

It is not difficult to prove that $A(R) = 0$ if the representation R is either real or pseudoreal. Indeed, if this is the case, then there is a matrix S (symmetric or antisymmetric) that intertwins the generators T_R^a and their complex conjugates $\overline{T}_R^a = -ST_R^aS^{-1}$. Then, using the hermiticity of the generators we can write

$$\mathrm{Tr}\,\left[T_R^a\{T_R^b, T_R^c\}\right] = \mathrm{Tr}\,\left[T_R^a\{T_R^b, T_R^c\}\right]^T = \mathrm{Tr}\,\left[\overline{T}_R^a\{\overline{T}_R^b, \overline{T}_R^c\}\right]. \tag{473}$$

[25] Schur's lemma states that if there is a matrix A that commutes with all elements of an irreducible representation of a Lie algebra, then $A = \lambda\mathbf{1}$, for some $\lambda \in \mathbb{C}$.

Now, using (465) or (466) we have

$$\text{Tr}\,\left[\overline{T}^a_R\{\overline{T}^b_R,\overline{T}^c_R\}\right] = -\text{Tr}\,\left[ST^a_RS^{-1}\{ST^b_RS^{-1},ST^c_RS^{-1}\}\right] = -\text{Tr}\,\left[T^a_R\{T^b_R,T^c_R\}\right], \qquad (474)$$

which proves that Tr $\left[T^a_R\{T^b_R,T^c_R\}\right]$ and therefore $A(R)=0$ whenever the representation is real or pseudoreal. Since the gauge anomaly in four dimensions is proportional to $A(R)$ this means that anomalies appear only when the fermions transform in a complex representation of the gauge group.

Acknowlegments

L.A.-G would like to thank the organizers of the 2012 Asia-Europe-Pacific School of High Energy Physics for their kindness and hospitality, and in particular Nick Ellis, Hellen Haller, Masami Yokoyama, Lydia Roos and Francesco Riva for plenty of interesting discussion. The work of M.A.V.-M. has been partially supported by Spanish Science Ministry Grants FPA2009-10612, FPA2012-34456, and FIS2012-30926, Basque Government Grant IT-357-07, and Spanish Consolider-Ingenio 2010 Programme CPAN (CSD2007-00042).

References

[1] L. Álvarez-Gaumé and M. A. Vázquez-Mozo, *An Invitation to Quantum Field Theory,* Springer 2011.

[2] J. D. Bjorken and S. D. Drell, *Relativistic Quantum Fields*, McGraw-Hill 1965.

[3] C. Itzykson and J.-B. Zuber, *Quantum Field Theory*, McGraw-Hill 1980.

[4] P. Ramond, *Field Theory: A Modern Primer*, Addison-Wesley 1990.

[5] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory*, Addison Wesley 1995.

[6] S. Weinberg, *The Quantum Theory of Fields*, Vols. 1-3, Cambridge 1995

[7] P. Deligne et al. (editors), *Quantum Fields and Strings: a Course for Mathematicians*, American Mathematical Society 1999.

[8] A. Zee, *Quantum Field Theory in a Nutshell*, Princeton 2003.

[9] B. S. DeWitt, *The Global Approach to Quantum Field Theory*, Vols. 1 & 2, Oxford 2003.

[10] V. P. Nair, *Quantum Field Theory. A Modern Perspective*, Springer 2005.

[11] T. Banks, *Modern Quantum Field Theory*, Cambridge 2008.

[12] O. Klein, *Die Reflexion von Elektronen an einem Potentialsprung nach der Relativischen Dynamik von Dirac*, Z. Phys. **53** (1929) 157.

[13] B. R. Holstein, *Klein's paradox*, Am. J. Phys. **66** (1998) 507.

[14] N. Dombey and A. Calogeracos, *Seventy years of the Klein paradox*, Phys. Rept. **315** (1999) 41.
N. Dombey and A. Calogeracos, *History and Physics of the Klein Paradox*, Contemp. Phys. **40** (1999) 313 (`quant-ph/9905076`).

[15] F. Sauter, *Zum Kleinschen Paradoxon*, Z. Phys. **73** (1932) 547.

[16] H. B. G. Casimir, *On the attraction between two perfectly conducting plates*, Proc. Kon. Ned. Akad. Wet. **60** (1948) 793.

[17] G. Plunien, B. Müller and W. Greiner, *The Casimir Effect*, Phys. Rept. **134** (1986) 87.
K. A. Milton, *The Casimir Effect: Physical Manifestation of Zero-Point Energy*, (`hep-th/9901011`).
K. A. Milton, *The Casimir effect: recent controversies and progress*, J. Phys. **A37** (2004) R209 (`hep-th/0406024`).
S. K. Lamoreaux, *The Casimir force: background, experiments, and applications*, Rep. Prog. Phys. **68** (2005) 201.

[18] M. J. Sparnaay, *Measurement of attractive forces between flat plates*, Physica **24** (1958) 751.

[19] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover 1972.

[20] Y. Aharonov and D. Bohm, *Significance of the electromagnetic potentials in the quantum theory*, Phys. Rev. **115** (1955) 485.

[21] P. A. M. Dirac, *Quantised Singularities in the Electromagnetic Field*, Proc. Roy. Soc. **133** (1931) 60.

[22] S. Dodelson, *Modern Cosmology*, Academic Press 2003.

[23] P. A. M. Dirac, *Lectures on Quantum Mechanics*, Dover 2001.

[24] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems*, Princeton 1992.

[25] R. Jackiw, *Quantum meaning of classical field theory*, Rev. Mod. Phys. **49** (1977) 681
R. Jackiw, *Introduction to the Yang-Mills quantum theory*, Rev. Mod. Phys. **52** (1980) 661.

[26] P. Ramond, *Journeys Beyond the Standard Model*, Perseus Books 1999.
R. .N . Mohapatra, *Unification and Supersymmetry. The Frontiers of Quark-Lepton Physics*, Springer 2003.

[27] C. P. Burguess, *Goldstone and pseudogoldstone bosons in nuclear, particle and condensed matter physics*, Phys. Rept. **330** (2000) 193 (`hep-th/9808176`).

[28] L. Álvarez-Gaumé, *An introduction to anomalies*, in: "Fundamental problems of gauge field theory", eds. G. Velo and A. S. Wightman, Plenum Press 1986.

[29] R. Jackiw, *Topological investigations of quantized gauge theories*, in: "Current Algebra and Anomalies", eds. S. B. Treiman, R. Jackiw, B. Zumino and E. Witten, Princeton 1985.

[30] S. Adler, *Axial-Vector Vertex in Spinor Electrodynamics,* Phys. Rev. **177** (1969) 2426.
J. S. Bell and R. Jackiw, *A PCAC puzzle:* $\pi^0 \to 2\gamma$ *in the sigma model*, Nuovo Cimento **A60** (1969) 47.

[31] J. Steinberger, *On the Use of Substraction Fiels and the Lifetimes of Some Types of Meson Decay*, Phys. Rev. **76** (1949) 1180.

[32] F. J. Ynduráin, *The Theory of Quark and Gluon Interactions*, Springer 1999.

[33] G. 't Hooft, *How the instantons solve the U(1) problem*, Phys. Rept. **142** (1986) 357.

[34] D. G. Sutherland, *Current Algebra and Some Nonstrong Mesonic Decays*, Nucl. Phys. **B2** (1967) 433.
M. J. G. Veltman, *Theoretical aspects of high-energy neutrino interactions*, Proc. R. Soc. **A301** (1967) 107.

[35] S. L. Adler and W. A. Bardeen, *Absence of higher order corrections in the anomalous axial vector divergence equation*, Phys. Rev. **182** (1969) 1517.

[36] E. Witten, *An SU(2) anomaly*, Phys. Lett. **B117** (1982) 324.

[37] S. Eidelman et al. *Review of Particle Phhysics*, Phys. Lett. **B592** (2004) 1 (`http://pdg.lbl.gov`).

[38] D. J. Gross and F. Wilczek, *Ultraviolet behavior of nonabelian gauge theories*, Phys. Rev. Lett. **30** (1973) 1343.

[39] H. D. Politzer, *Reliable perturbative results for strong interations?*, Phys. Rev. Lett. **30** (1973) 1346.

[40] G. 't Hooft, remarks at the *Colloquium on Renormalization of Yang-Mills fields and applications to particle physics*, Marseille 1972.

[41] I. B. Khriplovich, *Green's functions in theories with a non-abelian gauge group*, Yad. Fiz. **10** (1969) 409 [Sov. J. Nucl. Phys. **10** (1970) 235].
M. V. Terentiev and V. S. Vanyashin, *The vacuum polarization of a charged vector field*, Zh. Eskp. Teor. Fiz. **48** (1965) 565 [Sov. Phys. JETP **21** (1965) 375].

[42] K. G. Wilson, *Renormalization group and critical phenomena 1. Renormalization group and the*

Kadanoff scaling picture, Phys. Rev. **B4** (1971) 3174.
K. G. Wilson, *Renormalization group and critical phenomena 2. Phase space cell analysis of critical behavior*, Phys. Rev. **B4** (1971) 3184
K. G. Wilson, *The renormalization group and critical phenomena*, Rev. Mod. Phys. **55** (1983) 583.

[43] L. P. Kadanoff, *Scaling Laws for Ising Models Near T_c*, Physics **2** (1966) 263.

[44] J. Schwinger, *On Gauge Invariance and Vacuum Polarization*, Phys. Rev. **82** (1951) 664.

[45] E. Brezin and C. Itzykson, *Pair Production in Vacuum by an Alternating Field*, Phys. Rev. **D2** (1970) 1191.

[46] S. W. Hawking, *Particle Creation by Black Holes*, Commun. Math. Phys. **43** (1975) 199.

[47] M. K. Parikh and F. Wilczek, *Hawking Radiation as Tunneling*, Phys. Rev. Lett. **85** (2000) 5042 (`hep-th/9907001`)

[48] Yu. A. Golfand and E. P. Likhtman, *Extension of the Algebra of Poincaré group generators and violations of P-invariance*, JETP Lett. **13** (1971) 323.
D. V. Volkov and V. P. Akulov, *Is the Neutrino a Goldstone Particle*, Phys. Lett. **B46** (1973) 109.
J. Wess and B. Zumino, *A Lagrangian Model Invariant under Supergauge Transformations*, Phys. Lett. **B49** (1974) 52.

[49] R. Haag, J. Łopuszański and M. Sohnius, *All possible generators of supersymmetries of the S-matrix*, Nucl. Phys. **B88** (1975) 257.

Quantum Chromodynamics

Hsiang-nan Li
Institute of Physics, Academia Sinica, Taipei, Taiwan 115, Republic of China

Abstract
I review the basics of perturbative QCD, including infrared divergences and safety, collinear and k_T factorization theorems, and various evolution equations and resummation techniques for single- and double-logarithmic corrections. I then elaborate its applications to studies of jet substructures and hadronic two-body heavy-quark decays.

1 Introduction

One of the important missions of the Large Hadron Collider (LHC) is to search for new physics beyond the standard model. The identification of new physics signals usually requires precise understanding of standard-model background, whose contributions mainly arise from quantum chromodynamics (QCD). Many theoretical approaches have been developed based on QCD, which are appropriate for studies of processes in different kinematic regions and involving different hadronic systems. The theoretical framework for high-energy hadron collisions is known as the perturbative QCD (pQCD). I will focus on pQCD below, introducing its fundamental ingredients and applications to LHC physics. Supplementary material can be found in [1].

The simple QCD Lagrangian reveals rich dynamics. It exhibits the confinement at low energy, which accounts for the existence of various hadronic bound states, such as pions, protons, B mesons, and etc.. This nonperturbative dynamics is manifested by infrared divergences in perturbative calculations of bound-state properties like parton distribution functions and fragmentation functions. On the other hand, the asymptotic freedom at high energy leads to a small coupling constant, that allows formulation of pQCD. Therefore, it is possible to test QCD in high-energy scattering, which is, however, nontrivial due to bound-state properties of involved hadrons. That is, high-energy QCD processes still involve both perturbative and nonperturbative dynamics. A sophisticated theoretical framework needs to be established in order to realize the goal of pQCD: it is the factorization theorem [2], in which infrared divergences are factorized out of a process, and the remaining piece goes to a hard kernel. The point is to prove the universality of the infrared divergences, namely, the independence of processes the same hadron participates in. Then the infrared divergences are absorbed into a parton distribution function (PDF) for the hadron, which just needs to be determined once, either from experimental data or by nonperturbative methods. The universality of a PDF guarantees the infrared finiteness of hard kernels for all processes involving the same hadron. Convoluting these hard kernels with the determined PDF, one can make predictions. In other words, the universality of a PDF warrants the predictive power of the factorization theorem.

Though infrared divergences are factorized into a PDF, the associated logarithmic terms may appear in a process, that is not fully inclusive. To improve perturbative expansion, these logarithmic corrections should be organized by evolution equations or resummation techniques. For the summation of different single logarithms, the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equation [3] and the Balitsky-Fadin-Kuraev-Lipatov (BFKL) equation [4] have been proposed. For different double logarithms, the threshold resummation [5–7] and the k_T resummation [8, 9] have been developed. Besides, an attempt has been made to combine the DGLAP and BFKL equations, leading to the Ciafaloni-Catani-Fiorani-Marchesini (CCFM) equation [10]. Similarly, the threshold and k_T resummations has been unified under the joint resummation [11, 12], which is applicable to processes in a wider kinematic range. A simple framework for understanding all the above evolution equations and resummation techniques will be provided.

After being equipped with the pQCD formalism, we are ready to learn its applications to various processes, for which I will introduce jet substructures and hadronic two-body heavy-quark decays. It will be demonstrated that jet substructures, information which is crucial for particle identification at the LHC and usually acquired from event generators [13], are actually calculable using the resummation technique. Among jet substructures investigated in the literature, the distribution in jet invariant mass and the energy profile within a jet cone will be elaborated. For the latter, it will be shown that the factorization theorem goes beyond the conventional naive factorization assumption [14], and provides valuable predictions for branching ratios and CP asymmetries of hadronic two-body heavy-quark decays, that can be confronted by LHCb data. Specifically, I will concentrate on three major approaches, the QCD-improved factorization [15], the perturbative QCD [16–19], and the soft-collinear-effective theory [20–23]. Some long-standing puzzles in B meson decays and their plausible resolutions are reviewed. For more details on this subject, refer to [24].

2 Factorization Theorem

The QCD lagrangian is written as

$$\mathcal{L}_{QCD} = \bar{\psi}(i \not{D}_a T_a - m)\psi - \frac{1}{4} F_a^{\mu\nu} F_{\mu\nu a}, \tag{1}$$

with the quark field ψ, the quark mass m, and the covariant derivative and the gauge field tensor

$$\begin{aligned} D_a^\mu &= \partial^\mu + igA_a^\mu, \\ F_a^{\mu\nu} &= \partial^\mu A_a^\nu - \partial^\nu A_a^\mu - g f_{abc} A_b^\mu A_c^\nu, \end{aligned} \tag{2}$$

respectively. The color matrices T_a and the structure constants f_{abc} obey

$$[T_a^{(F)}, T_b^{(F)}] = i f_{abc} T_c^{(F)}, \quad (T_a^{(A)})_{bc} = -i f_{abc}, \tag{3}$$

where F (A) denotes the fundamental (adjoint) representation. Adding the gauge-fixing term in the path-integral quantization to remove spurious degrees of freedom, Eq. (1) becomes

$$\mathcal{L}_{QCD} = \bar{\psi}(i \not{D}_a T_a - m)\psi - \frac{1}{4} F_a^{\mu\nu} F_{\mu\nu a} - \frac{1}{2}\lambda(\partial_\mu A_a^\mu)^2 + \partial_\mu \eta_a^\dagger (\partial^\mu + g f_{abc} A_c^\mu)\eta_b, \tag{4}$$

with the gauge parameter λ, and the ghost field η. The last term in the above expression comes from the Jacobian for the variable change, as fixing the gauge.

The Feynman rules for QCD can be derived from Eq. (4) following the standard procedures [25]. The quark and gluon propagators with the momentum p are given by $i \not{p}/(p^2 + i\epsilon)$ and $-ig^{\mu\nu}/p^2$ in the Feynman gauge, respectively. The quark-gluon-quark vertex and the ghost-gluon-ghost vertex are written as $-ig\gamma_\mu T_a$ and $g f_{abc} p'_\mu$, respectively, where the subscripts μ and a are associated with the gluon, p' is the momentum of the outgoing ghost, and b (c) is associated with the outgoing (incoming) ghost. The three-gluon vertex and the four-gluon vertex are given by

$$\begin{aligned} \Gamma_{3g} &= -g f_{a_1 a_2 a_3}[g^{\nu_1\nu_2}(p_1 - p_2)^{\nu_3} + g^{\nu_2\nu_3}(p_2 - p_3)^{\nu_1} + g^{\nu_3\nu_1}(p_3 - p_1)^{\nu_2}, \\ \Gamma_{4g} &= -ig^2[f_{ea_1a_2}f_{ea_3a_4}(g^{\nu_1\nu_3}g^{\nu_2\nu_4} - g^{\nu_1\nu_4}g^{\nu_2\nu_3}) + f_{ea_1a_3}f_{ea_4a_2}(g^{\nu_1\nu_4}g^{\nu_3\nu_2} - g^{\nu_1\nu_2}g^{\nu_3\nu_4}) \\ &\quad + f_{ea_1a_4}f_{ea_2a_3}(g^{\nu_1\nu_2}g^{\nu_4\nu_3} - g^{\nu_1\nu_3}g^{\nu_4\nu_2})], \end{aligned} \tag{5}$$

respectively, where the subscripts a_1, a_2, $\cdots$ and ν_1, ν_2, $\cdots$ are assigned to gluons counterclockwise. The particle momenta flow into the vertices in all the above Feynman rules.

2.1 Infrared Divergences and Safety

The first step to establish the factorization theorem is to identify infrared divergences in Feynman diagrams for a QCD process at quark-gluon level. We start with the vertex correction to the amplitude $\gamma^*(q) \to q(p_1)\bar{q}(p_2)$, in which a virtual photon of momentum $q = p_1 + p_1$ splits into a quark of momentum p_1 and an anti-quark of momentum p_2. Given the Feynman rules, one has the loop integral

$$\int \frac{d^4l}{(2\pi)^4}(-ig\gamma^\nu T_a)\frac{i(\not{p}_1 - \not{l})}{(p_1-l)^2+i\epsilon}(-ie\gamma_\mu)\frac{-i(\not{p}_2 - \not{l})}{(p_2-l)^2+i\epsilon}(-ig\gamma_\nu T_a)\frac{-i}{l^2+i\epsilon}, \tag{6}$$

where l is the loop momentum carried by the gluon, and the inclusion of the corresponding counterterm for the regularization of a ultraviolet divergence is understood. The appearance of infrared divergences becomes more transparent, as performing the contour integration in the light-cone frame, in which the coordinates $l^\mu = (l^+, l^-, \mathbf{l}_T)$ are defined by

$$l^\pm = \frac{l^0 \pm l^z}{\sqrt{2}}, \quad \mathbf{l}_T = (l^x, l^y). \tag{7}$$

When an on-shell particle moves along the light cone, only one component of its momentum is large in this frame. For example, the above quark momenta can be chosen as $p_1^\mu = (p_1^+, 0, \mathbf{0}_T)$ and $p_2^\mu = (0, p_2^-, \mathbf{0}_T)$.

In terms of the light-cone coordinates, Eq. (6) is reexpressed as

$$\int \frac{dl^+dl^-d^2l_T}{(2\pi)^4}\frac{1}{2(l^+-p_1^+)l^- - l_T^2 + i\epsilon}\frac{1}{2l^+(l^- - p_2^-) - l_T^2 + i\epsilon}\frac{1}{2l^+l^- - l_T^2 + i\epsilon}, \tag{8}$$

where only the denominators are shown, since infrared divergences are mainly determined by pole structures. The poles of l^- are located, for $0 < l^+ < p_1^+$, at

$$l^- = \frac{l_T^2}{2(l^+ - p_1^+)} + i\epsilon, \;\; l^- = p_2^- + \frac{l_T^2}{2l^+} - i\epsilon, \;\; l^- = \frac{l_T^2}{2l^+} - i\epsilon. \tag{9}$$

As $l^+ \sim O(p_1^+)$, the contour of l^- is pinched at $l^- \sim O(l_T^2/p_1^+)$ by the first and third poles, defining the collinear region. As $l^+ \sim O(l_T)$, the contour of l^- is pinched at $l^- \sim O(l_T)$, defining the soft region. That is, the collinear (soft) region corresponds to the configuration of $l^\mu \sim (E, \Lambda^2/E, \Lambda)$ ($l^\mu \sim (\Lambda, \Lambda, \Lambda)$), where E and Λ denote a large scale and a small scale, respectively. Another leading configuration arises from the hard region characterized by $l^\mu \sim (E, E, E)$. A simple power counting implies that all the above three regions give logarithmic divergences. Picking up the first pole in Eq. (9), Eq. (8) becomes

$$\frac{-i}{2p_1^+}\int \frac{dl^+d^2l_T}{(2\pi)^3}\frac{p_1^+ - l^+}{2p_2^- l^+(p_1^+ - l^+) + p_1^+ l_T^2}\frac{1}{l_T^2} \approx \frac{-i}{4p_1\cdot p_2}\frac{1}{(2\pi)^3}\int \frac{dl^+}{l^+}\int \frac{d^2l_T}{l_T^2}, \tag{10}$$

which produces the double logarithm from the overlap of the collinear (the integration over l^+) and soft (the integration over l_T) enhancements.

The existence of infrared divergences is a general feature of QCD corrections. An amplitude is not a physical quantity, but a cross section is. To examine whether the infrared divergences really call for attention, we extend the calculation to the cross section of the process $e^-e^+ \to X$, the e^-e^+ annihilation into hadrons. A cross section is computed as the square of an amplitude, whose Feynman diagrams are composed of those for the amplitude connected by their complex conjugate with a final-state cut between them. The cross section at the Born level $e^-e^+ \to \gamma^* \to q\bar{q}$ is written as

$$\sigma^{(0)} = N_c\frac{4\pi\alpha^2}{3Q^2}\sum_f Q_f^2, \tag{11}$$

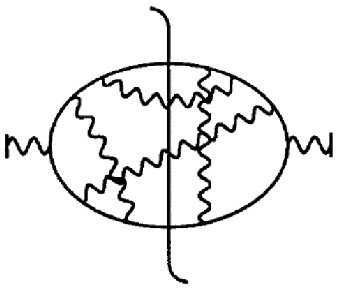


Fig. 1: Final-state cut on self-energy corrections to a virtual photon propagator.

where $N_c = 3$ is the number of colors, α is the electromagnetic coupling constant, Q^2 is the e^-e^+ invariant mass, and Q_f is the quark charge in units of the electron charge. The virtual one-loop corrections, including those to the gluon vertex in Eq. (6) and to the quark self-energy, give in the dimensional regularization [25]

$$\sigma^{(1)V} = -2N_c C_F \sum_f Q_f^2 \frac{\alpha\alpha_s}{\pi} Q^2 \left(\frac{4\pi\mu^2}{Q^2}\right)^{2\epsilon} \frac{1-\epsilon}{\Gamma(2-2\epsilon)} \left[\frac{1}{\epsilon^2} + \frac{3}{2}\frac{1}{\epsilon} - \frac{\pi^2}{2} + 4 + O(\epsilon)\right], \tag{12}$$

with the color factor $C_F = 4/3$, the strong coupling constant α_s, the renormalization scale μ, and the Gamma function Γ. The double pole $1/\epsilon^2$ is a consequence of the overlap of the collinear and soft divergences. The one-loop corrections from real gluons lead to [25]

$$\sigma^{(1)R} = 2N_c C_F \sum_f Q_f^2 \frac{\alpha\alpha_s}{\pi} Q^2 \left(\frac{4\pi\mu^2}{Q^2}\right)^{2\epsilon} \frac{1-\epsilon}{\Gamma(2-2\epsilon)} \left[\frac{1}{\epsilon^2} + \frac{3}{2}\frac{1}{\epsilon} - \frac{\pi^2}{2} + \frac{19}{4} + O(\epsilon)\right]. \tag{13}$$

It is a crucial observation that the infrared divergences cancel in the summation over the virtual and real corrections in Eqs. (12) and (13), respectively: the double and single poles have a minus sign in the former, but a plus sign in the latter. It is easy to understand the infrared cancellation by means of self-energy corrections to the propagator of a virtual photon. Since a virtual photon does not involve a low characteristic scale, the loop corrections must be infrared finite. As taking the final-state cut shown in Fig. 1, the imaginary piece of a particle propagator is picked up, $\mathrm{Im}(1/(p^2 + i\epsilon)) \propto \delta(p^2)$, which corresponds to the Feynman rule for an on-shell particle. Because the self-energy corrections are infrared finite, their imaginary part, i.e., the $e^-e^+ \to X$ cross section, is certainly infrared finite. The above observation has been formulated into the Kinoshita-Lee-Nauenberg (KLN) theorem [26], which states that a cross section is infrared safe, as integrating over all phase spaces of final states. Combining Eqs. (11), (12), and (13), one derives the $e^-e^+ \to X$ cross section up to next-to-leading order (NLO)

$$\sigma = N_c \frac{4\pi\alpha^2}{3Q^2} \sum_f Q_f^2 \left[1 + \frac{3}{4}\frac{\alpha_s(Q)}{\pi} C_F\right], \tag{14}$$

that has been used to determine the strong coupling constant $\alpha_s(Q)$ at the scale Q.

2.2 DIS and Collinear Factorization

Though a naive perturbation theory applies to the e^-e^+ annihilation, it fails for more complicated ones, such as the deeply inelastic scattering (DIS) of a nucleon by a lepton, $\ell(k)N(p) \to \ell(k') + X$. Even as the momentum transfer squared $-q^2 = (k-k')^2 \equiv Q^2$ is large, the quark-level cross section for the DIS suffers infrared divergences at high orders, which reflect the nonperturbative dynamics in the nucleon. A special treatment of the infrared divergences is then required. It will be demonstrated that they can be factorized out of the scattering process, and absorbed into a nucleon PDF.

Consider the two structure functions $F_{1,2}(x, Q^2)$ involved in the DIS, where the Bjorken variable is defined as $x \equiv -q^2/(2p \cdot q) = Q^2/(2p \cdot q)$, and take F_2 as an example. We shall not repeat loop integrations, but quote the NLO corrections to the quark-level diagrams [25]:

$$\begin{aligned} F_2^q(x, Q^2) &= x\left\{\delta(1-x) + \frac{\alpha_s}{2\pi} C_F \left[\frac{1+x^2}{1-x}\left(\ln\frac{1-x}{x} - \frac{3}{4}\right) + \frac{1}{4}(9+5x)\right]_+ \right. \\ &\left. + \frac{\alpha_s}{2\pi} C_F \left(\frac{1+x^2}{1-x}\right)_+ \left(4\pi\mu e^{-\gamma_E}\right)^\epsilon \int_0^{Q^2} \frac{dk_T^2}{k_T^{2+2\epsilon}} + \cdots \right\}, \end{aligned} \tag{15}$$

where the superscript q denotes the initial-state quark, γ_E is the Euler constant, and the first term comes from the leading-order (LO) contribution. The subscript $+$ represents the plus function, which is understood as a distribution function via

$$\int_0^1 dx \frac{f(x)}{(1-x)_+} \equiv \int_0^1 dx \frac{f(x) - f(1)}{1-x}. \tag{16}$$

The integration over k_T^2 generates an infrared divergence, that is regularized in the dimensional regularization with $\epsilon < 0$,

$$\int_0^{Q^2} \frac{dk_T^2}{k_T^{2+2\epsilon}} = \frac{1}{-\epsilon}\left(Q^2\right)^{-\epsilon}. \tag{17}$$

Hence, the infrared divergence does exist in the perturbative evaluation of the DIS structure function, even after summing over the virtual and real corrections. This divergence arises from the collinear region with the loop momentum being parallel to the nucleon momentum, since it can also be regularized by introducing a mass to the initial-state quark. It is related to the confinement mechanism, and corresponds to a long-distance phenomenon associated with a group of collimated on-shell particles. The other terms in Eq. (15) represent the hard NLO contribution to the structure function. Comparing the results for the DIS and for the e^-e^+ annihilation, the former involves the integration over final-state kinematics, but not over initial-state kinematics. This is the reason why the KLN theorem does not apply to the infrared divergences associated with the initial-state nucleon, and the above collinear divergence exists. Note that the soft divergences cancel between virtual and real diagrams due to the fact that a nucleon is color-singlet: a soft gluon with a huge space-time distribution cannot resolve the color structure of a nucleon, so it does not interact with it.

Besides, the collinear gluon emissions modify a quark momentum, such that the initial-state quark can carry various momenta, as it participates in hard scattering. It is then natural to absorb the collinear divergences into a PDF for the nucleon, $\phi_{q/N}$, which describes the probability for quark q to carry certain amount of the nucleon momentum. In other words, the quark-level collinear divergences are subtracted by those in the PDF in perturbation theory, and the remaining infrared finite piece contributes to the hard kernel H. We write the quark-level structure function as the following expansion in the strong coupling constant,

$$F_2^q(x, Q^2) = H^{(0)} \otimes \phi_{f/N}^{(0)} + \frac{\alpha_s}{2\pi} H^{(1)} \otimes \phi_{q/N}^{(0)} + \frac{\alpha_s}{2\pi} H^{(0)} \otimes \phi_{q/N}^{(1)} + \cdots, \tag{18}$$

where $H^{(i)}$ ($\phi_{q/N}^{(i)}$) is the hard kernel (PDF) of the i-th order. The symbol $\otimes$ represents a convolution in the parton momentum fraction ξ:

$$H \otimes \phi_{q/N} \equiv \int_x^1 \frac{d\xi}{\xi} H(x/\xi, Q, \mu)\phi_{q/N}(\xi, \mu). \tag{19}$$

We are ready to assign each term in Eq. (15) into either $H^{(i)}$ or $\phi^{(i)}_{q/N}$. The first term $\delta(1-x)$ goes to $H^{(0)}\otimes\phi^{(0)}_{q/N}$ with the definitions

$$H^{(0)}(x/\xi,Q,\mu)=\delta(1-x/\xi),\quad \phi^{(0)}_{q/N}(\xi,\mu)=\delta(1-\xi), \tag{20}$$

which confirm $H^{(0)}\otimes\phi^{(0)}_{q/N}=\delta(1-x)$. The second term in Eq. (15) is assigned to $H^{(1)}\otimes\phi^{(0)}_{q/N}$ and the third term to $H^{(0)}\otimes\phi^{(1)}_{q/N}$ with

$$\begin{aligned} H^{(1)}(x,Q,\mu) &= P^{(1)}_{qq}(x)\ln\frac{Q^2}{\mu^2}+\cdots, \\ \phi^{(1)}_{q/N}(\xi,\mu) &= \left(4\pi\mu e^{-\gamma}\right)^{\epsilon}P^{(1)}_{qq}(\xi)\int_0^{\mu^2}\frac{dk_T^2}{k_T^{2+2\epsilon}}, \end{aligned} \tag{21}$$

and the quark splitting function

$$P^{(1)}_{qq}(x)=C_F\left(\frac{1+x^2}{1-x}\right)_+. \tag{22}$$

The definition of the PDF in terms of a hadronic matrix element is given by

$$\begin{aligned} \phi_{q/N}(\xi,\mu) &= \int\frac{dy^-}{2\pi}\exp(-i\xi p^+y^-) \\ &\quad\times\frac{1}{2}\sum_\sigma\langle N(p,\sigma)|\bar q(0,y^-,0_T)\frac{1}{2}\gamma^+W(y^-,0)q(0,0,0_T)|N(p,\sigma)\rangle, \end{aligned} \tag{23}$$

where $|N(p,\sigma)\rangle$ denotes the bound state of the nucleon with momentum p and spin σ, y^- is the minus component of the coordinate of the quark field after the final-state cut, the first factor $1/2$ is attributed to the average over the nucleon spin, and the matrix $\gamma^+/2$ is the spin projector for the nucleon. Here μ is called the factorization scale, which is similar to a renormalization scale, but introduced in perturbative computations for an effective theory. The Wilson lines are defined by $W(y^-,0)=W(0)W^\dagger(y^-)$ with

$$W(y^-)=\mathcal{P}\exp\left[-ig\int_0^\infty dzn_-\cdot A(y+zn_-)\right], \tag{24}$$

where $\mathcal{P}$ represents a path-ordered exponential. The Wilson line behaves like a scalar particle carrying a color source. The two quark fields in Eq. (23) are separated by a distance, so the above Wilson links are demanded by the gauge invariance of the nonlocal matrix element. Since Eq. (23) depends only on the property of the nucleon, but not on the hard processes it participates in, a PDF is universal (process-independent). This is the most important observation, that warrants the predictive power of the factorization theorem.

The Wilson line appears as a consequence of the eikonalization of the final-state quark, to which the collinear gluons attach. The eikonalization is illustrated below by considering the loop correction to the virtual photon vertex. Assuming the initial-state quark momentum $p=(p^+,0,\mathbf{0_T})$ and the final-state quark momentum $p'=(0,p'^-,\mathbf{0_T})$, we have the partial integrand

$$\not{p}'\gamma^\nu\frac{\not{p}'+\not{l}}{(p'+l)^2}\gamma^\mu\frac{\not{p}+\not{l}}{(p+l)^2}\gamma_\nu\approx\not{p}'\gamma^-\frac{\not{p}'+\not{l}}{(p'+l)^2}\gamma^\mu\frac{\not{p}+\not{l}}{(p+l)^2}\gamma^+\approx\not{p}'\gamma^-\frac{\not{p}'}{2p'\cdot l}\gamma^\mu\frac{\not{p}+\not{l}}{(p+l)^2}\gamma^+, \tag{25}$$

as the loop momentum l is collinear to p, where $\not{p}'$ comes from the Feynman rule for the final-state quark, γ^μ is the photon vertex, and the subleading contribution from the transverse components of γ^ν has been

neglected. Applying the identity $\gamma^- \not{p}' = 2p'^- - \not{p}'\gamma^-$ and $\not{p}' \not{p}' = p'^2 = 0$ leads the above expression to

$$\not{p}'\gamma^\mu \frac{\not{p}+\not{l}}{(p+l)^2}\gamma^+ \frac{p'^-}{p'\cdot l} \approx \not{p}'\gamma^\mu \frac{\not{p}+\not{l}}{(p+l)^2}\gamma^+ \frac{n_-^-}{n_-\cdot l} \approx \not{p}'\gamma^\mu \frac{\not{p}+\not{l}}{(p+l)^2}\gamma_\nu \frac{n_-^\nu}{n_-\cdot l}, \tag{26}$$

where the dimensionless vector $n_- = (0, 1, \mathbf{0_T})$ is parallel to p', and the subleading contribution from $\nu = T$ has been restored. The factor n_-^ν and $1/n_-\cdot l$ are called the eikonal vertex and the eikonal propagator, respectively.

It is then shown that the Feynman rule $n_-^\nu/n_-\cdot l$ for the eikonalized final-state quark is derived from the Wilson line in Eq. (24). Consider the expansion of the path-order exponential in $W(0)$ up to order of α_s, and Fourier transform the gauge field into the momentum space,

$$\begin{aligned} & -ig\int_0^\infty dz n_-\cdot \int d^4l \exp[iz(n_-\cdot l + i\epsilon)]\tilde{A}(l) \\ = \; & -ig\int d^4l \frac{\exp[iz(n_-\cdot l+i\epsilon)]}{i(n_-\cdot l+i\epsilon)}\bigg|_{z=0}^{z=\infty} n_-\cdot\tilde{A}(l) = \int d^4l \frac{gn_-^\nu}{n_-\cdot l+i\epsilon}\tilde{A}_\nu(l), \end{aligned} \tag{27}$$

where the term $i\epsilon$ has been introduced to suppress the contribution from $z=\infty$. The field $\tilde{A}(l)$ is contracted with the gauge field from the initial-state quark with interaction to form the gluon propagator $-i/(l^2+i\epsilon)$. The expansion of the second piece $W(y^-)$ gives the Feynman rules for the eikonal propagator appearing after the final-state cut. In this case the additional exponential factor $\exp(il\cdot y)$ is combined with $\exp(-i\xi p^+ y^-)$, implying that the valence quark $q(0, y^-, 0_T)$ after the final-state cut carries the momentum $\xi p - l$. In summary, the first (second) piece of Wilson lines corresponds to the configuration without (with) the loop momentum flowing through the hard kernel. The above discussion verifies the Wilson lines in the PDF definition.

After detaching the collinear gluons from the final-state quark, the fermion flow still connects the PDF and the hard kernel. To achieve the factorization in the fermion flow, we insert the Fierz identity,

$$\begin{aligned} I_{ij}I_{lk} \;=\; & \frac{1}{4}I_{ik}I_{lj} + \frac{1}{4}(\gamma_\alpha)_{ik}(\gamma^\alpha)_{lj} + \frac{1}{4}(\gamma^5\gamma_\alpha)_{ik}(\gamma^\alpha\gamma^5)_{lj} \\ & + \frac{1}{4}(\gamma^5)_{ik}(\gamma^5)_{lj} + \frac{1}{8}(\gamma^5\sigma_{\alpha\beta})_{ik}(\sigma^{\alpha\beta}\gamma^5)_{lj}, \end{aligned} \tag{28}$$

with I being the identity matrix and $\sigma_{\alpha\beta} \equiv i[\gamma_\alpha, \gamma_\beta]/2$. At leading power, only the term $(\gamma_\alpha)_{ik}(\gamma^\alpha)_{lj}/4$ contributes, in which the structure $(\gamma^\alpha)_{lj}/2 \approx (\gamma^+)_{lj}/2$ goes to the definition of the PDF in Eq. (23), and $(\gamma_\alpha)_{ik}/2 \approx (\gamma^-)_{ik}/2$ goes into the evaluation of the hard kernel. The other terms in Eq. (28) contribute at higher powers. Similarly, we have to factorize the color flow between the PDF and the hard kernel by inserting the identity

$$I_{ij}I_{lk} = \frac{1}{N_c}I_{ik}I_{lj} + 2(T^c)_{ik}(T^c)_{lj}, \tag{29}$$

where I denotes the 3×3 identity matrix, and T^c is a color matrix. The first term in the above expression contributes to the present configuration, in which the valence quarks before and after the final-state cut are in the color-singlet state. The structure I_{lj}/N_c goes into the definition of the PDF, and I_{ik} goes into the evaluation of the hard kernel. The second term in Eq. (29) contributes to the color-octet state of the valence quarks, together with which an additional gluonic parton comes out of the nucleon and participates in the hard scattering.

The factorization formula for the nucleon DIS structure function is written as

$$F_2(x, Q^2) = \sum_f \int_x^1 \frac{d\xi}{\xi} H_f(x/\xi, Q, \mu)\phi_{f/N}(\xi, \mu), \tag{30}$$

with the subscript f labeling the parton flavor, such as a valence quark, a gluon, or a sea quark. The hard kernel H_f is obtained following the subtraction procedure for the collinear divergences, and its LO and NLO expressions have been presented in Eqs. (20) and (21), respectively. The universal PDF $\phi_{f/N}$, describing the probability for parton f to carry the momentum fraction ξ in the nucleon, takes a smooth model function. It must be derived by nonperturbative methods, or extracted from data.

2.3 Predictive Power

The factorization theorem derived above is consistent with the well-known parton model. The nucleon travels a long space-time, before it is hit by the virtual photon. As $Q^2 \gg 1$, the hard scattering occurs at point space-time. Relatively speaking, the quark in the nucleon behaves like a free particle before the hard scattering, and decouples from the rest of the nucleon. Therefore, the cross section for the nucleon DIS reduces to an incoherent sum over parton flavors under the collinear factorization. That is, the approximation

$$\left|\sum_i \mathcal{M}_{i/N}\right|^2 \approx \sum_i |\mathcal{M}_f|^2 \, \phi_{f/N}, \tag{31}$$

holds, where $\mathcal{M}_{i/N}$ represents the scattering amplitude for partonic state i of the nucleon N (it could be a multi-parton state), and $\mathcal{M}_f$ represents the infrared finite scattering amplitude for parton f.

Comparing the factorization theorem with the operator product expansion (OPE), the latter involves an expansion in short distance $y^\mu \sim 0$. A typical example is the infrared safe $e^- e^+ \to X$, whose cross section can be expressed as a series $\sigma \approx \sum_i C_i(y) O_i(0)$. The Wilson coefficients C_i and the local effective operators O_i appear in a product in the OPE. A factorization formula involves an expansion on the light cone with small $y^2 \sim 0$, instead of $y^\mu \sim 0$. A typical example is the DIS structure function, in which the existence of the collinear divergences implies that a parton travels a finite longitudinal distance y^-. It is also the reason why the hard kernel H_f and the PDF $\phi_{f/N}$ appear in a convolution in the momentum fraction.

The factorization procedure introduces the factorization scale μ into the hard kernel H_f and the PDF $\phi_{f/N}$, as indicated in Eq. (30). Higher-order corrections produce the logarithms $\ln(Q/\mu)$ in H_f and $\ln(\mu/Q_0)$ in $\phi_{f/N}$, which come from the splitting of $\ln(Q/Q_0)$ in the structure function F_2, Q_0 being a low scale characterizing $\phi_{f/N}$. One usually sets $\mu = Q$ to eliminate the logarithm in H_f, such that the input $\phi_{f/N}(\xi, Q)$ for arbitrary Q is needed. The factorization scale does not exist in QCD diagrams, but is introduced when a physical quantity like the structure function is factorized. The independence of the factorization scale, $\mu dF_2/d\mu = 0$, leads to a set of renormalization-group (RG) equations

$$\begin{aligned} \mu \frac{d}{d\mu} \phi_{f/N}(\xi, \mu) &= \gamma_f \phi_{f/N}(\xi, \mu), \\ \mu \frac{d}{d\mu} H_f(x/\xi, Q, \mu) &= -\gamma_f H_f(x/\xi, Q, \mu), \end{aligned} \tag{32}$$

where γ_f denotes the anomalous dimension of the PDF. A solution of the RG equations describes the evolution of the PDF in Q

$$\phi_{f/N}(\xi, Q) = \phi_{f/N}(\xi, Q_0) \exp\left[\int_{Q_0}^{Q} \frac{d\mu}{\mu} \gamma_f(\alpha_s(\mu))\right], \tag{33}$$

as a result of the all-order summation of $\ln(Q/Q_0)$. Hence, one just extracts the initial condition $\phi(\xi, Q_0)$ defined at the initial scale Q_0 from data. The PDF at other higher scales Q is known through the evolution. That is, the inclusion of the RG evolution increases the predictive power of the factorization theorem.

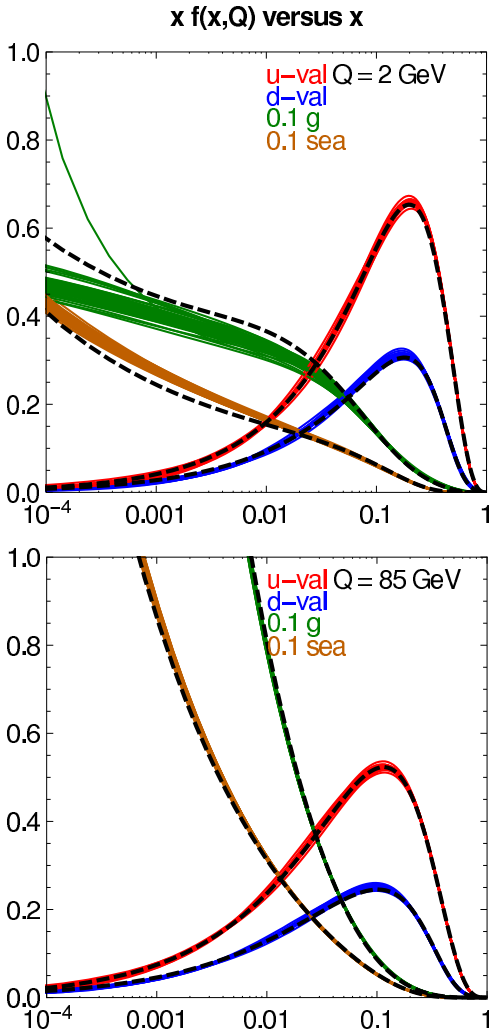


Fig. 2: CT10 NNLO (solid color) and NLO (dashed) parton distribution functions.

Fitting the factorization formulas for those processes, whose dynamics is believed to be clear, such as Eq. (30) for DIS, one has determined the PDFs for various partons in the proton. The CTEQ-TEA CT10 models at the accuracy of NLO and next-to-next-to-leading order (NNLO) for hard kernels are displayed in Figs. 2 [27, 28]. The increase of the gluon and sea-quark PDFs with the decrease of the momentum fraction ξ is a consequence of more radiations in that region in order to reach a lower ξ. The comparison of the PDFs at $Q = 2$ GeV and $Q = 85$ GeV indicates that the valence u-quark and d-quark PDFs become broader with Q, while the gluon and sea-quark PDFs increase with Q.

Note that a choice of an infrared regulator is, like an ultraviolet regulator, arbitrary; namely, we can associate an arbitrary finite piece with the infrared pole $1/(-\epsilon)$ in $\phi^{(1)}_{f/N}$. Shifts of different finite pieces between $\phi_{f/N}$ and H_f correspond to different factorization schemes. Hence, the extraction of a PDF depends not only on powers and orders, at which QCD diagrams are computed, but on factorization schemes. Since perturbative calculations are performed up to finite powers and orders, a factorization scheme dependence is unavoidable. Nevertheless, the scheme dependence of pQCD predictions would be minimized, if one sticks to the same factorization scheme. Before adopting models for PDFs, it should be checked at which power and order, at which initial scale, and in what scheme they are determined.

At last, I explain how to apply the factorization theorem to make predictions for QCD processes. A nucleon PDF $\phi_{f/N}$ is infrared divergent, if evaluated in perturbation theory due to the confinement mechanism. The QCD diagram for a DIS structure function involving quarks and gluons as the external

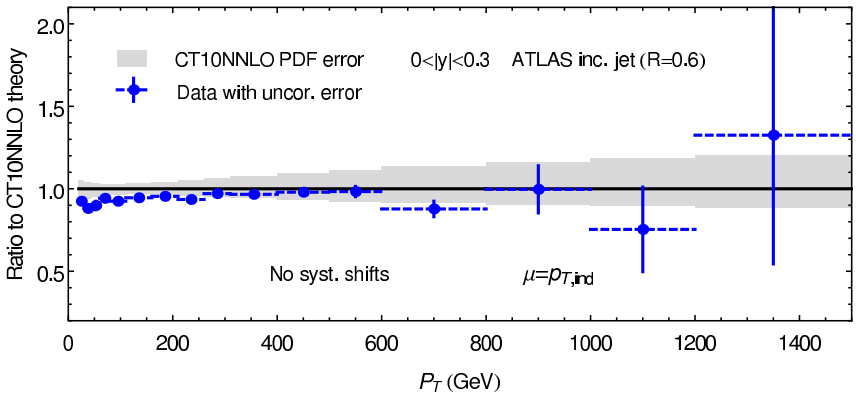


Fig. 3: Comparison of ATLAS data for inclusive jet p_T distribution with a theoretical prediction using CT10 NNLO.

particles are also infrared divergent. It has been demonstrated that the infrared divergences cancel between the QCD diagrams and the effective diagrams for $\phi_{f/N}$, as taking their difference, which defines the hard kernel $H^{\rm DIS}$. One then derives the factorization formula for other processes, such as the Drell-Yan (DY) process $N(p_1)N(p_2) \to \ell^+\ell^-(q) + X$, and computes the corresponding hard kernel $H^{\rm DY}$. The point is to verify that the infrared divergences in the QCD diagrams for DY and in the effective diagrams for the nucleon PDF cancel, and $H^{\rm DY}$ is infrared finite. If it is the case, the universality of the nucleon PDF holds, and the factorization theorem is applicable. If not, the factorization theorem fails. After verifying the factorization theorem, one makes predictions for the DY cross section using the formula $\sigma^{\rm DY} = \phi_{f_1/N} \otimes H^{\rm DY} \otimes \phi_{f_2/N}$. As an example, the predictions for the inclusive jet p_T distribution derived from the factorization theorem [28] are presented in Fig. 3. The consistency between the predictions and the ATLAS data is obvious.

2.4 k_T Factorization

The collinear factorization theorem introduced above has been intensively investigated and widely applied to many QCD processes up to higher powers and orders. The evolution of PDFs from low to high factorization scales is governed by the DGLAP equation. The databases for PDFs have been constructed, such as the CTEQ models. Other nonperturbative inputs like soft functions, jet functions, and fragmentation functions have been all explored to some extent. However, another more complicated framework, the k_T factorization theorem [29–31], may be more appropriate in some kinematic regions or in semi-inclusive processes. The collinear factorization applies, when the DIS is measured at a finite Bjorken variable x. The cross section is written as the convolution of a hard kernel with a PDF in a parton momentum fraction ξ. As $x \to 0$, $\xi \geq x$ can reach a small value, at which the parton transverse momentum k_T is of the same order of magnitude as the longitudinal momentum ξp, and not negligible. Once k_T is kept in a hard kernel, a transverse-momentum-dependent (TMD) function $\Phi(\xi, k_T, \mu)$ is needed to describe the parton distribution not only in the momentum fraction ξ, but also in the transverse momentum k_T. The DIS cross section is then written, in the k_T factorization theorem, as the convolution

$$F_2(x, Q^2) = \sum_f \int_x^1 \frac{d\xi}{\xi} \int d^2k_T H_f(x/\xi, k_T, Q, \mu)\Phi_{f/N}(\xi, k_T, \mu). \tag{34}$$

The k_T factorization theorem is also applicable to the analysis of low p_T spectra of final states, like direct photon and jet productions, for which $k_T \sim p_T$ is not negligible.

A collinear gluon emission, modifying a parton longitudinal momentum, generates a parton transverse momentum k_T at the same time. The factorization of a TMD from the DIS is similar to that of a PDF, which relies on the eikonal approximation in the collinear region. This procedure results in the

eikonal propagator $n_-^\nu/n_-\cdot l$, represented by the Wilson lines similar to that defined in Eq. (24). A naive TMD definition as an extension of the PDF in Eq. (23) is given by

$$\Phi_{q/N}(\xi,k_T,\mu) = \int\frac{dy^-}{2\pi}\int\frac{d^2y_T}{(2\pi)^2}e^{-i\xi p^+y^-+i\mathbf{k}_T\cdot\mathbf{y}_T} \times\frac{1}{2}\langle N(p,\sigma)|\bar{q}(0,y^-,y_T)\frac{1}{2}\gamma^+W(y^-,y_T,0,0_T)q(0,0,0_T)|N(p,\sigma)\rangle, \quad (35)$$

with the Wilson links $W(y^-,y_T,0,0_T)=W(0,0_T)I_{0,y_T}W^\dagger(y^-,y_T)$. Because the valence quark fields before and after the final-state cut are separated by a transverse distance in this case, the vertical links I_{0,y_T} located at $y^-=\infty$ are demanded by the gauge invariance of a TMD [32]. More investigations on the vertical Wilson links can be found in [33].

Though we do need the k_T factorization theorem, many of its aspects have not yet been completely understood. For example, the naive definition in Eq. (35) is actually ill-defined, due to the existence of the light-cone singularity, that arises from a loop momentum parallel to the Wilson line direction n_-. A plausible modification is to rotate the Wilson line away from the light cone, namely, to replace n_- by a vector n with $n^2\neq 0$. This rotation is allowed, since the collinear divergences are insensitive to the direction n as illustrated in Eq. (26) [34]: even when n_- is rotated to n, only the minus component n^- is relevant for the evaluation of the collinear divergences. A detailed discussion on this subtle issue can be found in [35]. Besides, a parton is off-shell by $-k_T^2$, once k_T is retained. Then whether a hard kernel obtained in the k_T factorization theorem is gauge invariant becomes a concern [36]. Dropping the k_T dependence of the hard kernel in Eq. (34), the integration of the TMD over k_T, $\int d^2k_T\Phi_{f/N}(\xi,k_T)$, can be worked out. How this integral is related to the PDF $\phi_{f/N}(\xi)$ in Eq. (23) is worth of a thorough study.

3 Evolution and resummation

As stated in the previous section, radiative corrections in pQCD produce large logarithms at each order of the coupling constant. Double logarithms appear in processes involving two scales, such as $\ln^2(p^+b)$ with p^+ being the large longitudinal momentum of a parton and $1/b$ being the small inverse impact parameter, where b is conjugate to the parton transverse momentum k_T. In the region with large Bjorken variable x, there exists $\ln^2(1/N)$ from the Mellin transformation of $\ln(1-x)/(1-x)_+$, for which the two scales are the large p^+ and the small infrared cutoff $(1-x)p^+$ for gluon emissions from a parton. Single logarithms are generated in processes involving one scale, such as $\ln p^+$ and $\ln(1/x)$, for which the relevant scales are the large p^+ and the small xp^+, respectively. Various methods have been developed to organize these logarithmic corrections to a PDF or a TMD: the k_T resummation for $\ln^2(p^+b)$ [8,9], the threshold resummation for $\ln^2(1/N)$ [5–7], the joint resummation [11,12] that unifies the above two formalisms, the DGLAP equation for $\ln p^+$ [3], the BFKL equation for $\ln(1/x)$ [4], and the CCFM equation [10] that combines the above two evolution equations. I will explain the basic ideas of all the single- and double-logarithmic summations in the Collins-Soper-Sterman (CSS) resummation formalism [8,9].

3.1 Resummation Formalism

Collinear and soft divergences may overlap to form double logarithms in extreme kinematic regions, such as low p_T and large x. The former includes low p_T jet, photon, and W boson productions, which all require real gluon emissions with small p_T. The latter includes top pair production, DIS, DY production, and heavy meson decays $B\to X_u l\nu$ and $B\to X_s\gamma$ [16, 37, 38] at the end points, for which parton momenta remain large, and radiations are constrained in the soft region. Because of the limited phase space for real gluon corrections, the infrared cancellation is not complete. The double logarithms, appearing in products with the coupling constant α_s, such as $\alpha_s\ln^2(E/p_T)$ with the beam energy E and $\alpha_s\ln(1-x)/(1-x)_+$, deteriorate perturbative expansion. Double logarithms also occur in exclusive

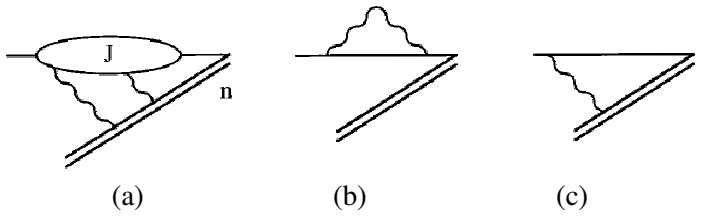


Fig. 4: (a) Jet subprocess defined in Eq. (36). (b) and (c) LO diagrams of (a).

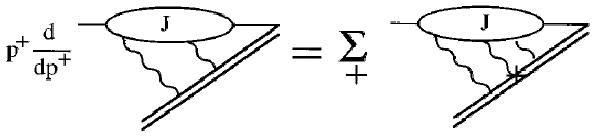


Fig. 5: Derivative $p^+ dJ/dp^+$ in the covariant gauge.

processes, such as Landshoff scattering [39], hadron form factors [40], Compton scattering [41] and heavy-to-light transitions $B \to \pi(\rho)$ [42] and $B \to D^{(*)}$ [43] at maximal recoil. In order to have a reliable pQCD analysis of these processes, the important logarithms must be summed to all orders.

The resummation of large logarithms will be demonstrated in the covariant gauge $\partial \cdot A = 0$ [38], in which the role of the Wilson line direction n and the key technique can be explained straightforwardly. Take as an example a jet subprocess defined by the matrix element

$$J(p,n)u(p) = \langle 0|\mathcal{P}\exp\left[-ig\int_0^\infty dz n\cdot A(nz)\right]q(0)|p\rangle\ , \tag{36}$$

where q is a light quark field with momentum p, and $u(p)$ is a spinor. The abelian case of this subprocess has been discussed in [44]. The path-ordered exponential in Eq. (36) is the consequence of the factorization of collinear gluons with momenta parallel to p from a full process, as explained in the previous section. For convenience, it is assumed that p has a large light-cone component p^+, and all its other components vanish. A general diagram of the jet function J is shown in Fig. 4(a), where the path-ordered exponential is represented by a double line along the vector n. As explained before, varying the direction n does not change the collinear divergences collected by the Wilson line.

It is easy to see that J contains double logarithms from the overlap of collinear and soft divergences by calculating the LO diagrams in Fig. 4(b), the self-energy correction, and in Fig. 4(c), the vertex correction. In the covariant gauge both Figs. 4(b) and 4(c) produce double logarithms. In the axial gauge $n \cdot A = 0$ the path-ordered exponential reduces to an identity, and Fig. 4(c) does not exist. The essential step in the resummation technique is to derive a differential equation $p^+ dJ/dp^+ = CJ$ [16, 38, 42], where the coefficient function C contains only single logarithms, and can be treated by RG methods. Since the path-ordered exponential is scale-invariant in n, J must depend on p and n through the ratio $(p\cdot n)^2/n^2$. The differential operator d/dp^+ can then be replaced by d/dn using a chain rule

$$p^+\frac{d}{dp^+}J = -\frac{n^2}{v\cdot n}v_\alpha\frac{d}{dn_\alpha}J, \tag{37}$$

with the vector $v = (1, 0, \mathbf{0}_T)$ being defined via $p = p^+ v$.

Equation (37) simplifies the analysis tremendously, because n appears only in the Feynman rules for the Wilson line, while p may flow through the whole diagram in Fig. 4(a). The differentiation of each eikonal vertex and of the associated eikonal propagator with respect to n_α,

$$-\frac{n^2}{v\cdot n}v_\alpha\frac{d}{dn_\alpha}\frac{n_\mu}{n\cdot l} = \frac{n^2}{v\cdot n}\left(\frac{v\cdot l}{n\cdot l}n_\mu - v_\mu\right)\frac{1}{n\cdot l} \equiv \frac{\hat{n}_\mu}{n\cdot l}, \tag{38}$$

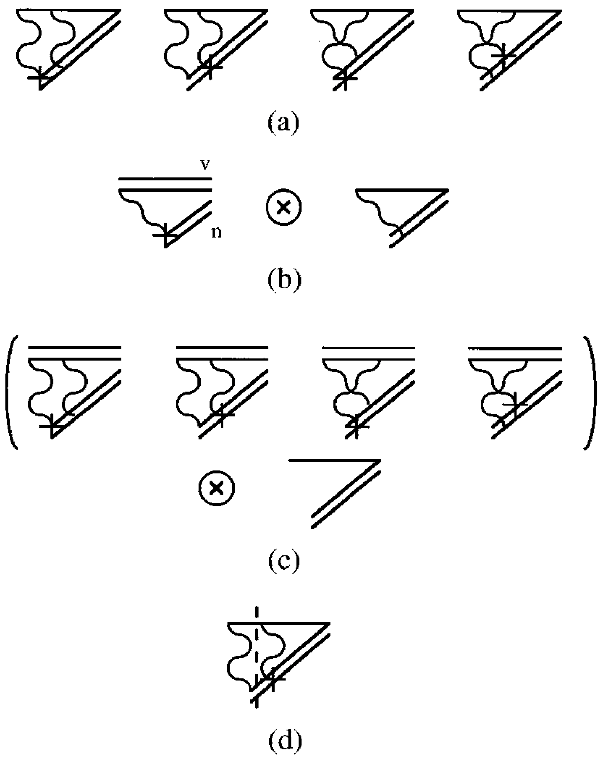


Fig. 6: (a) $O(\alpha_s^2)$ examples for the differentiated J. (b) Factorization of K at $O(\alpha_s)$. (c) Factorization of K at $O(\alpha_s^2)$. (d) Factorization of G at $O(\alpha_s)$.

leads to the special vertex $\hat{n}_\mu$. The derivative $p^+ dJ/dp^+$ is thus expressed as a summation over different attachments of $\hat{n}_\mu$, labeled by the symbol $+$ in Fig. 5. If the loop momentum l is parallel to p, the factor $v \cdot l$ vanishes, and $\hat{n}_\mu$ is proportional to v_μ. When this $\hat{n}_\mu$ is contracted with a vertex in J, in which all momenta are mainly parallel to p, the contribution to $p^+ dJ/dp^+$ is suppressed. Therefore, the leading regions of l are soft and hard.

According to this observation, we investigate some two-loop examples exhibited in Fig. 6(a). If the loop momentum flowing through the special vertex is soft but another is not, only the first diagram is important, giving a large single logarithm. In this soft region the subdiagram containing the special vertex can be factorized using the eikonal approximation as shown in Fig. 6(b), where the symbol $\otimes$ represents a convoluting relation. The subdiagram is absorbed into a soft kernel K, and the remainder is identified as the original jet function J, both being $O(\alpha_s)$ contributions. If both the loop momenta are soft, the four diagrams in Fig. 6(a) are equally important. The subdiagrams, factorized according to Fig. 6(c), contribute to K at $O(\alpha_s^2)$, and the remainder is the LO diagram of J. If the loop momentum flowing through the special vertex is hard and another is not, the second diagram in Fig. 6(a) dominates. In this region the subdiagram containing the special vertex is factorized as shown in Fig. 6(d). The right-hand side of the dashed line is absorbed into a hard kernel G as an $O(\alpha_s)$ contribution, and the left-hand side is identified as the $O(\alpha_s)$ diagram of J. If both the loop momenta are hard, all the diagrams in Fig. 6(a) are absorbed into G, giving the $O(\alpha_s^2)$ contributions.

Extending the above reasoning to all orders, one derives the differential equation

$$p^+ \frac{d}{dp^+} J = \left[K(m/\mu, \alpha_s(\mu)) + G(p^+ \nu/\mu, \alpha_s(\mu)) \right] J, \tag{39}$$

where the coefficient function C has been written as the sum of the soft kernel K and the hard kernel G. In the above expression μ is a factorization scale, the gauge factor in G is defined as $\nu = \sqrt{(v \cdot n)^2/|n^2|}$,

and a gluon mass m has been introduced to regularize the infrared divergence in K. It has been made explicit that K and G depend on a single infrared scale m and a single ultraviolet scale p^+, respectively.

The $O(\alpha_s)$ contribution to K from Fig. 6(b) is written as

$$K = -ig^2 C_F \mu^\epsilon \int \frac{d^{4-\epsilon}l}{(2\pi)^{4-\epsilon}} \frac{\hat{n}_\mu}{n\cdot l} \frac{g^{\mu\nu}}{l^2-m^2} \frac{v_\nu}{v\cdot l} - \delta K, \tag{40}$$

δK being an additive counterterm. The $O(\alpha_s)$ contribution to G from Fig. 6(d) is given by

$$G = -ig^2 C_F \mu^\epsilon \int \frac{d^{4-\epsilon}l}{(2\pi)^{4-\epsilon}} \frac{\hat{n}_\mu}{n\cdot l} \frac{g^{\mu\nu}}{l^2} \left(\frac{\not{p}+\not{l}}{(p+l)^2}\gamma_\nu - \frac{v_\nu}{v\cdot l} \right) - \delta G, \tag{41}$$

where the second term in the parentheses acts as a soft subtraction to avoid double counting, and δG is an additive counterterm. A straightforward evaluation shows that Eqs. (40) and (41) contain only the single logarithms $\ln(m/\mu)$ and $\ln(p^+\nu/\mu)$, respectively, as claimed before. Organizing these single logarithms using RG methods, and then solving Eq. (39), one resums the double logarithms $\ln^2(p^+/m)$ in J.

To explain all the known resummations and evolution equations, we first construct a master equation for the TMD $\Phi(x, k_T)$, which is a differential equation in the hadron momentum p^+. The dependence on the factorization scale μ is implicit. If the parton is a quark, Φ is defined by Eq. (35). If the parton is a gluon, the nonlocal operator in the hadronic matrix element of Eq. (35) is replaced by $F^+_\mu(y^-, y_T)F^{\mu+}(0)$. Similarly, n is varied arbitrarily away from the light cone with $n^2 \neq 0$. Then Φ depends on p^+ via the ratio $(p\cdot n)^2/n^2$, so the chain rule in Eq. (37) relating the derivative $d\Phi/dp^+$ to $d\Phi/dn_\alpha$ applies. Following the derivation in the previous subsection, one obtains the master equation

$$p^+ \frac{d}{dp^+}\Phi(x, k_T) = 2\bar{\Phi}(x, k_T), \tag{42}$$

where $\bar{\Phi}$ contains the special vertex, and the coefficient 2 is attributed to the equality of $\bar{\Phi}$ with the special vertex on either side of the final-state cut.

The function $\bar{\Phi}$ is factorized into the convolution of the soft and hard kernels with Φ:

$$\bar{\Phi}(x, k_T) = \bar{\Phi}_s(x, k_T) + \bar{\Phi}_h(x, k_T), \tag{43}$$

with the soft contribution

$$\begin{aligned} \bar{\Phi}_s &= \left[-ig^2 C_F \mu^\epsilon \int \frac{d^{4-\epsilon}l}{(2\pi)^{4-\epsilon}} \frac{\hat{n}\cdot v}{n\cdot l l^2 v\cdot l} - \delta K \right] \Phi(x, k_T) \\ &\quad -ig^2 C_F \mu^\epsilon \int \frac{d^{4-\epsilon}l}{(2\pi)^{4-\epsilon}} \frac{\hat{n}\cdot v}{n\cdot l v\cdot l} 2\pi i\delta(l^2)\Phi(x + l^+/p^+, |\mathbf{k}_T + \mathbf{l}_T|), \end{aligned} \tag{44}$$

where the first term is the same as in Eq. (40), and the second term proportional to $\delta(l^2)$ arises from the real soft gluon emission. The hard contribution is given by $\bar{\Phi}_h(x, k_T) = G(xp^+\nu/\mu, \alpha_s(\mu))\Phi(x, k_T)$, in which the hard kernel G is the same as in Eq. (41).

3.2 k_T Resummation and BFKL Equation

The TMD definition in Eq. (35) contains three scales: $(1-x)p^+$, xp^+, and k_T. We first consider the soft approximation corresponding to the rapidity ordering of real gluon emissions in a ladder diagram. Assume that a parton carries the longitudinal momentum $xp^+ + l_2^+ + l_1^+$, which becomes $xp^+ + l_1^+$ after emitting a gluon of longitudinal momentum l_2^+ and transverse momentum l_{2T}, and then becomes xp^+ after emitting a gluon of longitudinal momentum l_1^+ and transverse momentum l_{1T}. In the kinematic configuration with $l_2^+ \gg l_1^+$ and $l_{2T} \sim l_{1T}$, the original parton momentum is approximated by $xp^+ + l_2^+ + l_1^+ \approx xp^+ + l_2^+$. The loop integral associated with the first gluon emission is then independent

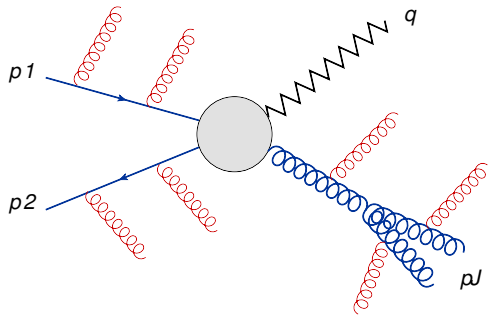


Fig. 7: Scattering amplitude for direct photon production.

of l_1^+, and can be worked out straightforwardly, giving a logarithm. The loop integral associated with the second gluon emission, involving only l_1^+, also gives a logarithm. Therefore, a ladder diagram with N rung gluons generates the logarithmic correction $(\alpha_s L)^N$ under the above rapidity ordering, where L denotes the large logarithm. Following the rapidity ordering, we adopt the approximation for the real gluon emission in Eq. (44)

$$\Phi(x + l^+/p^+, |\mathbf{k_T} + \mathbf{l_T}|) \approx \Phi(x, |\mathbf{k_T} + \mathbf{l_T}|), \tag{45}$$

where the l^+ dependence has been neglected. The transverse momenta l_T, being of the same order as k_T in this kinematic configuration, is kept. The variable l^+ in K is then integrated up to infinity, such that the scale $(1-x)p^+$ disappears.

Equation (44) is Fourier transformed into the impact parameter b space to decouple the l_T integration. Hence, in the intermediate x region Φ involves two scales, the large xp^+ that characterizes the hard kernel G and the small $1/b$ that characterizes the soft kernel K. The master equation (42) becomes

$$p^+ \frac{d}{dp^+}\Phi(x,b) = 2\left[K(1/(b\mu),\alpha_s(\mu)) + G(xp^+\nu/\mu,\alpha_s(\mu))\right]\Phi(x,b), \tag{46}$$

whose solution with $\nu = 1$ leads to the k_T resummation

$$\Phi(x,b) = \Delta_k(x,b)\Phi_i(x), \tag{47}$$

with the Sudakov exponential

$$\Delta_k(x,b) = \exp\left[-2\int_{1/b}^{xp^+}\frac{dp}{p}\int_{1/b}^{p}\frac{d\mu}{\mu}\gamma_K(\alpha_s(\mu))\right], \tag{48}$$

and the initial condition Φ_i of the Sudakov evolution. The anomalous dimension of K, $\lambda_K = \mu d\delta K/d\mu$, is given, up to two loops, by [45]

$$\gamma_K = \frac{\alpha_s}{\pi}C_F + \left(\frac{\alpha_s}{\pi}\right)^2 C_F\left[C_A\left(\frac{67}{36} - \frac{\pi^2}{12}\right) - \frac{5}{18}n_f\right], \tag{49}$$

with n_f being the number of quark flavors and $C_A = 3$ being a color factor.

The k_T resummation effect on the low p_T spectra of the direct photon production depicted in Fig. (7)has been analyzed [46]. The initial-state and final-state radiations are constrained in the low p_T region, where the k_T resummation is necessary for improving the perturbation theory. Figure 8 shows the deviation (Data -Theory)/Theory of the NLO pQCD predictions, obtained using the CTEQ4M PDFs [47], from the experimental data as a function of $x_t = 2p_T/\sqrt{s}$, $\sqrt{s}$ being the center-of-mass energy. The deviation is huge as expected, especially at low x_t of each set of the data. After including

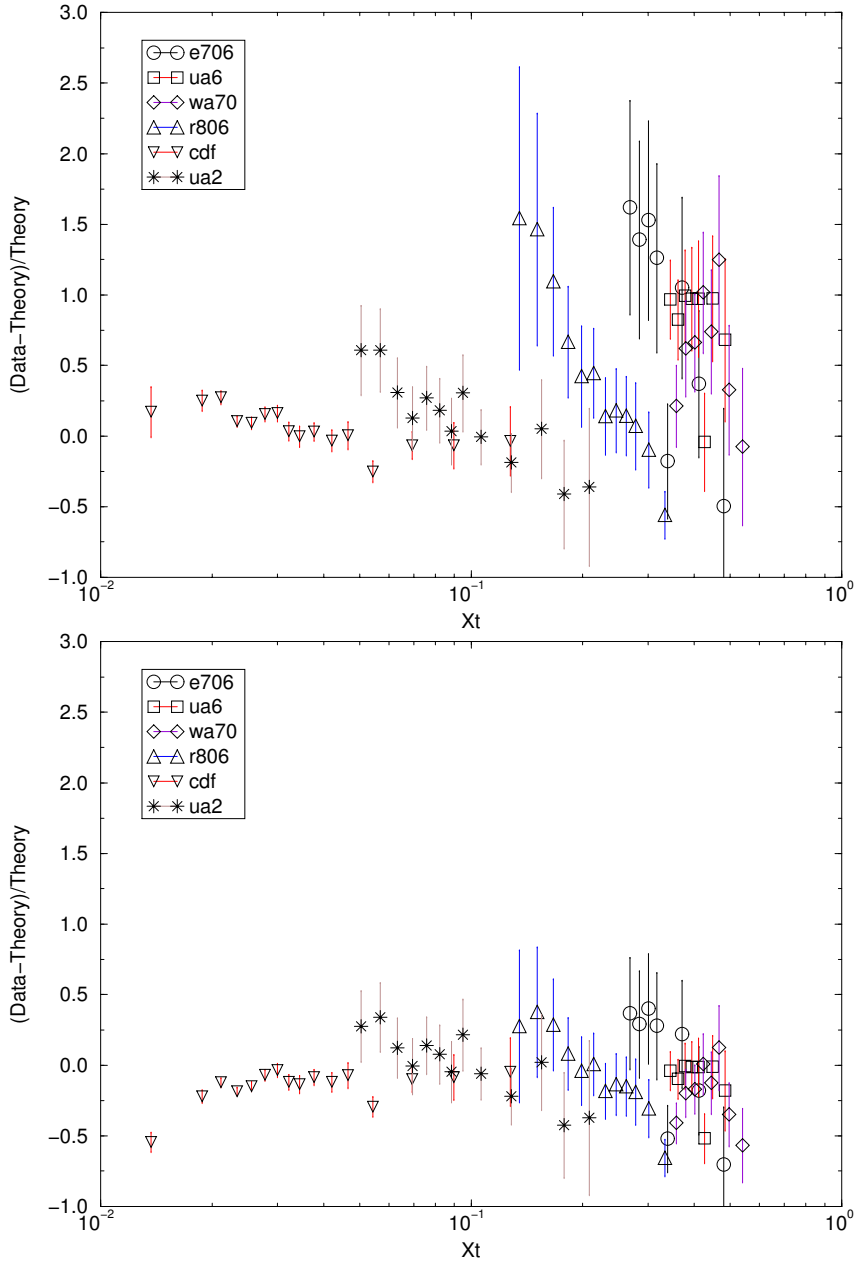


Fig. 8: Low p_T direct photon spectra before (upper) and after (lower) including the k_T resummation.

the k_T resummation effect [46], it is clear that a significant improvement on the agreement between theoretical predictions and the data is achieved. As to the intermediate- and high-p_T regions of the direct photon production, NLO pQCD works reasonably well in accommodating the data as indicated in Fig. 9. The threshold resummation effect, which will be introduced in the next subsection, is more relevant in these regions: it slightly improves the consistency between predictions and the data [48].

In the small x region with $xp^+ \sim k_T$, or $xp^+ \sim 1/b$ in the b space, the two-scale case reduces to the single-scale one. In this region contributions from gluonic partons dominate, so Φ represents the gluon TMD below. The source of double logarithms, i.e., the integral containing the anomalous dimension γ_K, is less important. Because only the soft scale exists, one drops the hard kernel G, and keeps the soft kernel with an ultraviolet cutoff. The right-hand side of Eq. (42) becomes

$$\bar{\Phi}(x,k_T) \;=\; -ig^2N_c\int\frac{d^4l}{(2\pi)^4}\frac{\hat{n}\cdot v}{n\cdot l v\cdot l}\left[\frac{\theta(k_T^2-l_T^2)}{l^2}\Phi(x,k_T)\right.$$

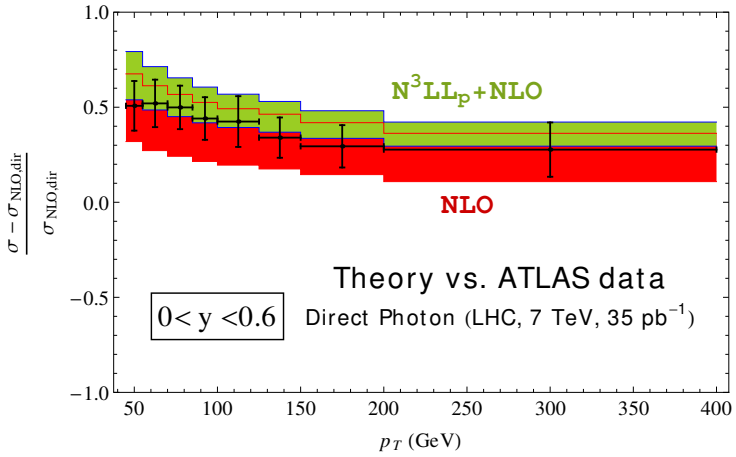


Fig. 9: High p_T direct photon spectrum under the threshold resummation.

$$+2\pi i\delta(l^2)\phi(x,|\mathbf{k}_T+\mathbf{l}_T|)\big]\,, \tag{50}$$

where the color factor C_F has been replaced by N_c for the gluon TMD. The θ function introduces the ultraviolet cutoff on l_T mentioned above. To make variation in x via variation in p^+, a fixed parton momentum is assumed. Under this assumption, the momentum fraction x is proportional to $1/p^+$, and one has $p^+ d\Phi/dp^+ = -x d\Phi/dx\Phi$ [49]. Performing the integrations over l^+ and l^- in Eq. (50), the master equation (42) reduces to the BFKL equation [50],

$$\frac{d\phi(x,k_T)}{d\ln(1/x)} = \bar{\alpha}_s \int \frac{d^2l_T}{\pi l_T^2}\left[\phi(x,|\mathbf{k}_T+\mathbf{l}_T|) - \theta(k_T^2-l_T^2)\phi(x,k_T)\right], \tag{51}$$

with the coupling constant $\bar{\alpha}_s = N_c\alpha_s/\pi$.

A remarkable prediction of the above LO BFKL equation is that a high-energy cross section increases with the center-of-mass energy,

$$\sigma \approx \frac{1}{t}\left(\frac{s}{t}\right)^{\omega_P - 1}, \tag{52}$$

with the momentum transfer squared t. It turns out that Eq. (52), with the Pomeron intercept $\omega_P - 1 = 4\bar{\alpha}_s \ln 2$, violates the Froissart (unitarity) bound $\sigma < \text{const.} \times \ln^2$ [51]. The unsatisfactory prediction of the LO BFKL equation called for the NLO corrections [52], which were, however, found to be dramatic as indicated by the x dependence of the derivative of the structure function $dF_L/d\ln Q^2$ in Fig. 10 [53]: the NLO effect is nearly as large as the LO result for $x \sim 0.001$, and becomes dominant at lower x. It even turns $dF_L/d\ln Q^2$ negative below $x \sim 0.0001$ in the upper of Fig. 10. That is, the perturbative solution is not at all stable. Choosing a running coupling constant [53], the NLO effect is not overwhelming, but still significant as exhibited in the lower of Fig. 10.

3.3 Threshold Resummation and DGLAP Equation

We then consider the soft approximation corresponding to the k_T ordering of real gluon emissions in a ladder diagram. Assume that a parton without the transverse momentum, carries $-\mathbf{l}_{1T}$ after emitting a gluon of longitudinal momentum l_1^+ and transverse momentum $\mathbf{l}_{1T}$, and then carries $-\mathbf{l}_{1T} - \mathbf{l}_{2T}$ after emitting a gluon of longitudinal momentum l_2^+ and transverse momentum $\mathbf{l}_{2T}$. In the kinematic configuration with $l_{2T} \gg l_{1T}$ and $l_2^+ \sim l_1^+$, the final parton momentum can be approximated by $-\mathbf{l}_{2T} - \mathbf{l}_{1T} \approx -\mathbf{l}_{2T}$, such that the loop integral associated with the first gluon emission involves only

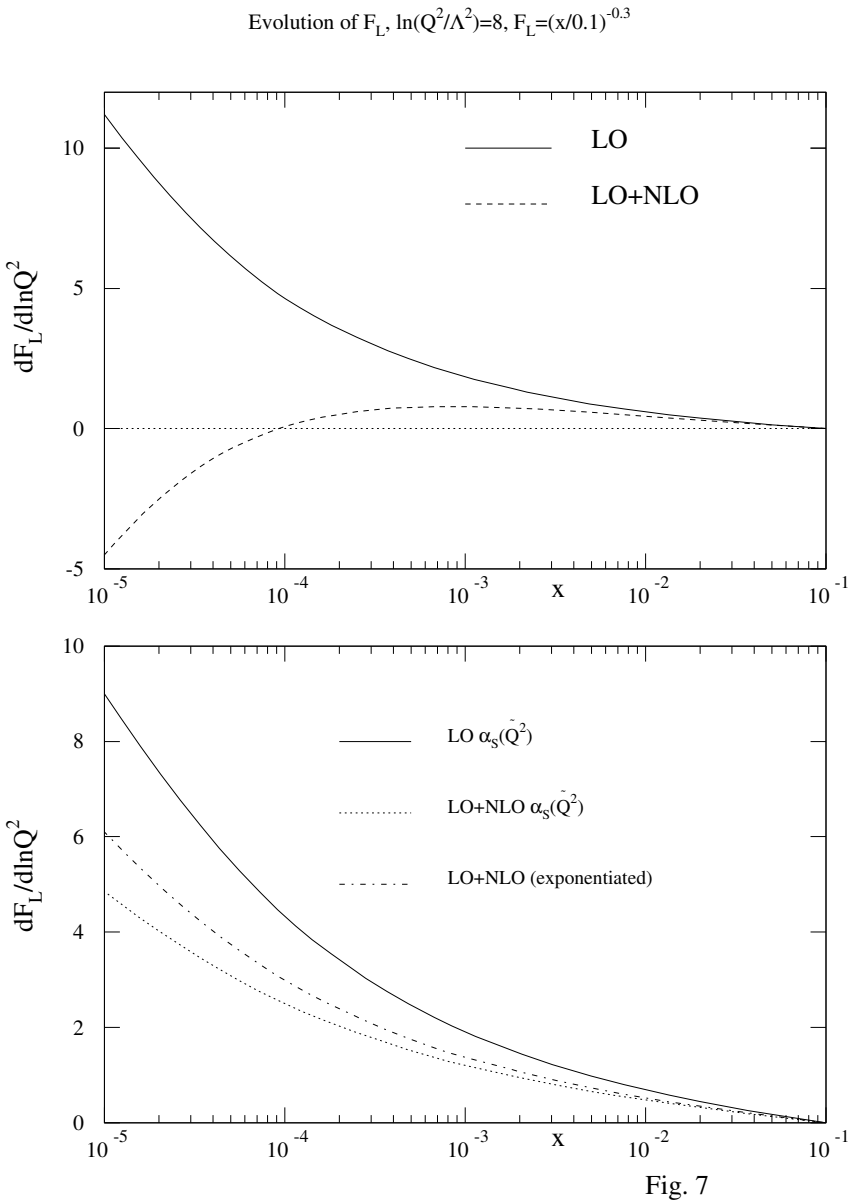


Fig. 10: Effects from LO and NLO BFKL equations.

l_{1T}, and can be worked out straightforwardly, giving a logarithm. The loop integral associated with the second gluon emission involves only l_{2T}, and also gives a logarithm. Hence, a ladder diagram with N rung gluons generates the logarithmic correction $(\alpha_s L)^N$ under the above k_T ordering. In this case Φ is independent of l_T, and we have the approximation for the real gluon emission in Eq. (44)

$$\Phi(x + l^+/p^+, |\mathbf{k_T} + \mathbf{l_T}|) \approx \Phi(x + l^+/p^+, k_T), \tag{53}$$

in which x and l^+/p^+ are of the same order. The dependence on k_T can then be integrated out from both sides of the master equation (42), and the TMD Φ reduces to the PDF ϕ. The scale k_T disappears, and the scale $(1-x)p^+$ is retained.

The Mellin transformation is employed to bring $\bar{\phi}_s$ from the momentum fraction x space to the moment N space,

$$\bar{\phi}_s(N) \;=\; \int_0^1 dx x^{N-1} \bar{\phi}_s(x), \tag{54}$$

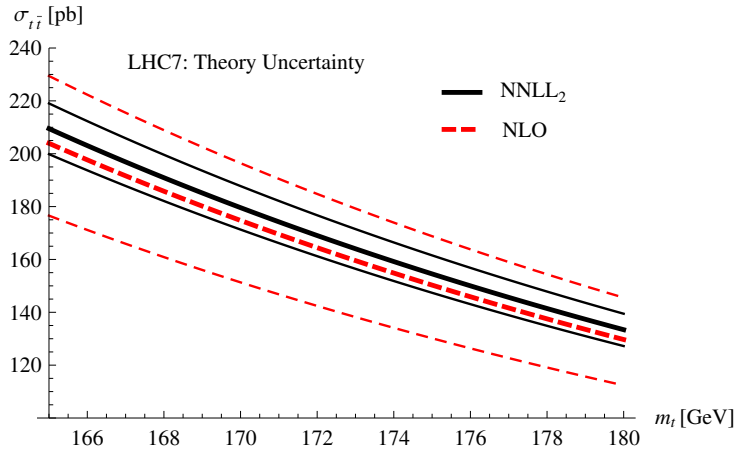


Fig. 11: Dependence of the total cross section for the top-pair production on the top mass at the LHC with $\sqrt{s} =$ 7 TeV.

under which the l^+ integration decouples. In the large x region ϕ involves two scales, the large $xp^+ \sim p^+$ from the hard kernel G and the small $(1-x)p^+ \sim p^+/N$ from the soft kernel K. To sum $\ln(1/N)$, we rewrite the derivative $p^+ d\phi/dp^+$ as

$$p^+ \frac{d\phi}{dp^+} = \frac{p^+}{N} \frac{d\phi}{d(p^+/N)}. \tag{55}$$

The solution of the master equation (42) then gives the threshold resummation,

$$\phi(N) = \Delta_t(N)\phi_i \tag{56}$$

with the exponential

$$\Delta_t(N) = \exp\left[-2\int_{p^+/N}^{p^+} \frac{dp}{p}\int_{p^+}^{p} \frac{d\mu}{\mu}\gamma_K(\alpha_s(\mu))\right], \tag{57}$$

or its equivalent expression

$$\Delta_t(N) = \exp\left[\int_0^1 dz \frac{1-z^{N-1}}{1-z}\int_{(1-z)^2}^1 \frac{d\lambda}{\lambda}\gamma_K(\alpha_s(\sqrt{\lambda}p^+))\right]. \tag{58}$$

An application of the threshold resummation is found in the analysis of the top-quark pair production, which was performed at the next-to-next-to-leading-logarithmic (NNLL) accuracy [54]. It has been observed that the threshold resummation effect enhances the NLO total cross section by few percents as shown in Fig. 11, where the bands sandwiched by the thinner lines denote the theory uncertainty. The above formalism can be used to determine of the top quark mass as indicated in Fig. 12, where the solid lines represent the central values, and the total uncertainties of the theoretical and experimental results [55] are given by the external dashed lines.

In the intermediate x region the two-scale case reduces to the single-scale one because of $xp^+ \sim (1-x)p^+$, and the source of double logarithms is less important. Without the Mellin transformation, the sum in Eq. (43), with the approximation in Eq. (53) being inserted, leads to the DGLAP equation [49],

$$p^+ \frac{d}{dp^+}\phi(x) = \int_x^1 \frac{d\xi}{\xi} P(x/\xi)\phi(\xi)\ , \tag{59}$$

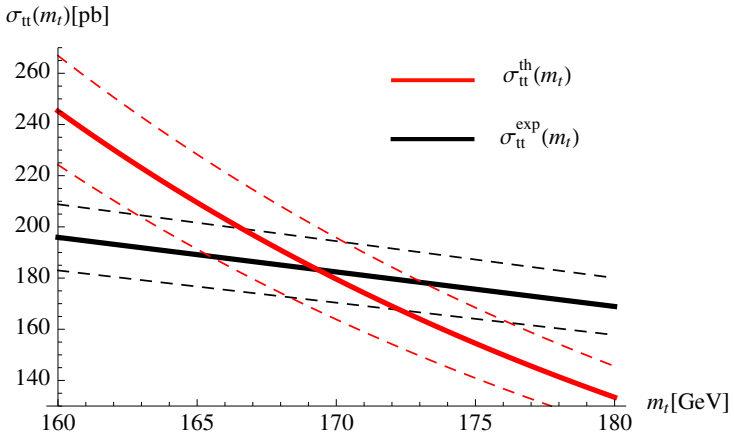


Fig. 12: Mass dependence of the theoretical cross section with the threshold resummation effect (red) and of the measured cross section (black).

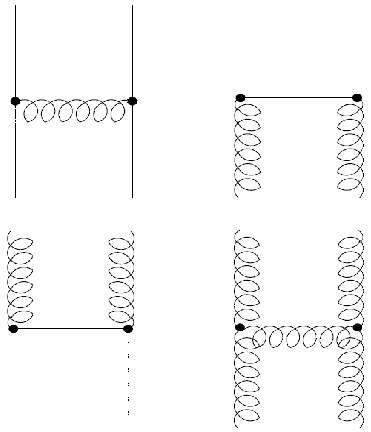


Fig. 13: Diagrams for the DGLAP splitting functions.

with the kernel

$$P(z) = \frac{\alpha_s(p^+)}{\pi} C_F \frac{2}{(1-z)_+} , \tag{60}$$

where the variable change $\xi = x + l^+/p^+$ has been made. The argument of α_s, i.e, the factorization scale μ, has been set to the scale $xp^+ \sim (1-x)p^+ \sim O(p^+)$. Note that the kernel P differs from the splitting function P_{qq} in Eq. (22) by the term $(z^2-1)/(1-z)_+$, which is finite in the $z \to 1$ limit. The reason is that the real gluon emission was evaluated under the soft approximation as deriving P, while it was calculated exactly as deriving P_{qq}.

Gluon emissions in Fig. 13 cause the mixing between the quark and gluon PDFs, giving the complete set of DGLAP equations with four splitting functions

$$\frac{\partial}{\partial \ln Q^2} \begin{pmatrix} \phi_q \\ \phi_g \end{pmatrix} = \begin{pmatrix} P_{qq} & P_{qg} \\ P_{gq} & P_{gg} \end{pmatrix} \otimes \begin{pmatrix} \phi_q \\ \phi_g \end{pmatrix} . \tag{61}$$

The evolution of the u-quark and d-quark PDFs in Q^2 predicted by the LO DGLAP equation [56] is shown in Fig. 14, where the inputs at the initial scale $Q_0 = 1$ GeV were taken from MRST2001 [57]. It is observed that the valence quark PDFs increase with Q^2 at small x, namely, they become broader with Q^2, a feature consistent with what was stated in the previous section. The predictions for the

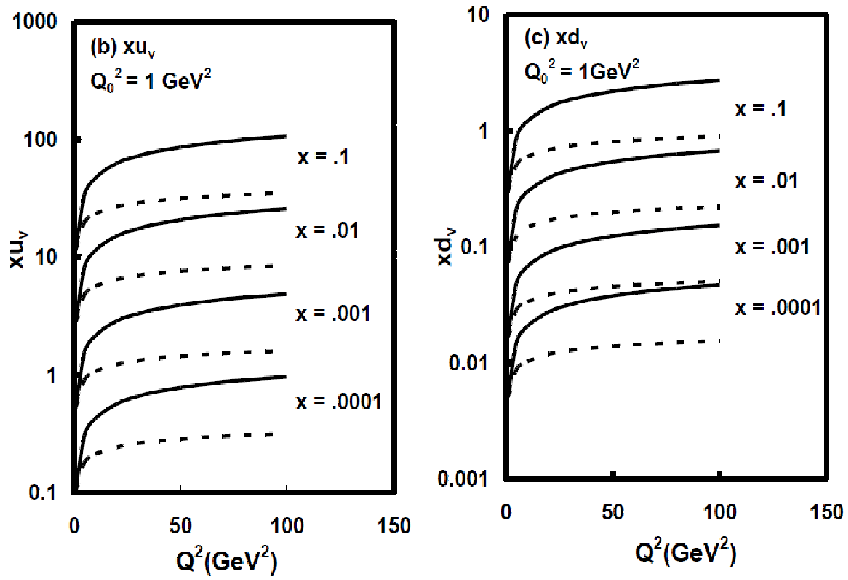


Fig. 14: Q^2 evolutions of the valence quark PDFs for some parameter values in the DGLAP solutions (solid and dashed lines).

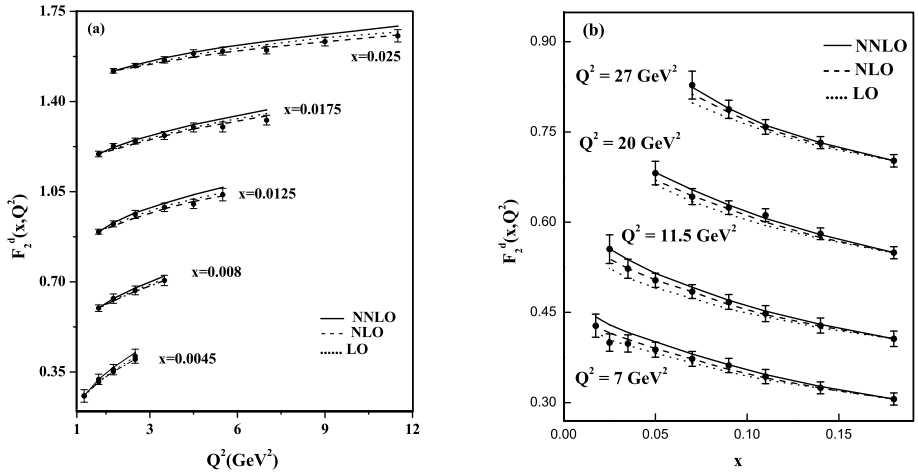


Fig. 15: Predictions from the DGLAP equation and the NMC data for the deuteron structure function. For clarity, data are scaled up by +0.2i (in Fig.(a)) and +i (in Fig.(b)) (with i = 0,1,2,3) starting from the bottom of the graphs.

deuteron structure function derived from the LO, NLO, and NNLO DGLAP equations are displayed in Fig. 15 [58], which agree with the NMC data [59].

3.4 Joint Resummation and CCFM Equation

At last, a unified resummation formalism for large and intermediate x and a unified evolution equation for intermediate and small x can be derived by retaining the l^+ and l_T dependencies of Φ in Eq. (44), which corresponds to the so-called angular ordering. In this case both the Fourier and Mellin transformations are applied to Eq. (44), leading to

$$\bar{\Phi}_s(N,b) = K(p^+/(N\mu), 1/(b\mu), \alpha_s(\mu))\Phi(N,b) \ , \tag{62}$$

with the soft kernel [11]

$$\begin{aligned} K &= -ig^2 C_F \mu^\epsilon \int_0^1 dz \int \frac{d^{4-\epsilon}l}{(2\pi)^{4-\epsilon}} \frac{\hat{n}\cdot v}{n\cdot l v\cdot l} \left[\frac{\delta(1-z)}{l^2}\right. \\ &\quad \left. +2\pi i\delta(l^2)\delta\left(1-z-\frac{l^+}{p^+}\right) z^{N-1} e^{i\mathbf{l}_T\cdot\mathbf{b}}\right] - \delta K, \\ &= \frac{\alpha_s(\mu)}{\pi} C_F \left[\ln\frac{1}{b\mu} - K_0\left(\frac{2\nu p^+ b}{N}\right)\right], \end{aligned} \tag{63}$$

K_0 being the modified Bessel function. As $p^+ b \gg N$, we have $K_0 \to 0$, and the soft scale inferred by the above expression approaches $1/b$ for the k_T resummation. As $N \gg p^+ b$, we have $K_0 \approx -\ln(\nu p^+ b/N)$, and the soft scale approaches p^+/N for the threshold resummation.

Following the procedures similar to Eqs. (46)-(48), we derive the joint resummation

$$\Phi(N,b) = \Delta_u(N,b)\Phi_i, \tag{64}$$

with the exponential

$$\Delta_u(N,b) = \exp\left[-2\int_{p^+\chi^{-1}(N,b)}^{p^+} \frac{dp}{p}\int_{p^+\chi^{-1}(1,b)}^{p} \frac{d\mu}{\mu}\gamma_K(\alpha_s(\mu))\right]. \tag{65}$$

The dimensionless function [12]

$$\chi(N,b) = \left(N + \frac{p^+ b}{2}\right) e^{\gamma_E}, \tag{66}$$

is motivated by the limits discussed above. It is apparent that Eq. (65) reduces to Eq. (48) and Eq. (57) in the $b \to \infty$ and $N \to \infty$ limits, respectively. The effect from the joint resummation on the q_T spectra of selectron pairs produced at the LHC with $\sqrt{S} = 14$ TeV has been investigated in [60]. It is seen in Fig. 16 that the joint and k_T resumations exhibit a similar behavior in the small-q_T region as expected, but the jointly-resummed cross section is about 5%-10% lower than the k_T-resummed cross section in the range 50 GeV $< q_T <$ 100 GeV.

In the intermediate and small x regions, it is not necessary to resum the double logarithms $\ln^2(1/N)$. After extracting the k_T resummation, the remaining single-logarithmic summation corresponds to a unification of the DGLAP and BFKL equations, since both the l^+ and l_T dependencies have been retained. The function $\Phi(x + l^+/p^+, b)$ in Eq. (44) is reexpressed, after the Fourier transformation, as

$$\begin{aligned} \Phi(x + l^+/p^+, b) &= \theta((1-x)p^+ - l^+)\Phi(x,b) \\ &\quad +[\Phi(x + l^+/p^+, b) - \theta((1-x)p^+ - l^+)\Phi(x,b)]. \end{aligned} \tag{67}$$

The contribution from the first term is combined with the first term in Eq. (44), giving the soft kernel K for the k_T resummation. The second term in Eq. (67) contributes

$$-iN_c g^2 \int \frac{d^4 l}{(2\pi)^4} \frac{\hat{n}\cdot v}{n\cdot l v\cdot l} 2\pi i\delta(l^2) e^{i\mathbf{l}_T\cdot\mathbf{b}} [\Phi(x + l^+/p^+, b) - \theta((1-x)p^+ - l^+)\Phi(x,b)], \tag{68}$$

which will generate the splitting function below. The color factor has been replaced by N_c, since the gluon TMD is considered here.

The master equation (42) then becomes

$$p^+ \frac{d}{dp^+}\Phi(x,b) = -2\left[\int_{1/b}^{xp^+} \frac{d\mu}{\mu}\gamma_K(\alpha_s(\mu)) - \bar{\alpha}_s(xp^+)\ln(p^+ b)\right]\Phi(x,b)$$

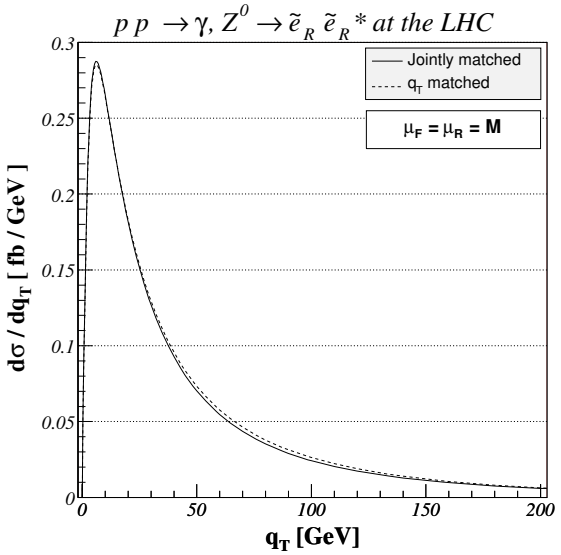


Fig. 16: Transverse-momentum distribution of selectron pairs at the LHC in the framework of joint (full) and k_T (dotted) resummations.

$$+2\bar{\alpha}_s(xp^+)\int_x^1 dz P_{gg}(z)\Phi(x/z,b), \tag{69}$$

with the splitting function

$$P_{gg}=\left[\frac{1}{(1-z)_+}+\frac{1}{z}-2+z(1-z)\right], \tag{70}$$

obtained from Eq. (68). The term $-2+z(1-z)$ finite as $z\to 0$ and $z\to 1$ has been added. The exponential Δ is extracted from the k_T resummation,

$$\Delta(x,b,Q_0)=\exp\left(-2\int_{xQ_0}^{xp^+}\frac{dp}{p}\left[\int_{1/b}^{p}\frac{d\mu}{\mu}\gamma_K(\alpha_s(\mu))-\bar{\alpha}_s(p)\ln\frac{pb}{x}\right]\right), \tag{71}$$

Q_0 being an arbitrary low energy scale. It is trivial to justify by substitution that the solution is given by

$$\begin{aligned}\Phi(x,b) &= \Delta(x,b,Q_0)\Phi_i \\ &\quad +2\int_x^1 dz\int_{Q_0}^{p^+}\frac{d\mu}{\mu}\bar{\alpha}_s(x\mu)\Delta_k(x,b)P_{gg}(z)\Phi(x/z,b),\end{aligned} \tag{72}$$

which can be regarded as a modified version of the CCFM equation [10].

4 PQCD for jet physics

Jets, abundantly produced at colliders [61], carry information of hard scattering and parent particles, which is crucial for particle identification and new physics search. Study of jet physics is usually done using event generators, which, however, suffer ambiguity from parameter tuning. Hence, we are motivated to establish an alternative approach free of the ambiguity. I will demonstrate that jet dynamics can be explored and jet properties can be predicted in the pQCD resummation formalism.

We start from the dijet production in the e^-e^+ annihilation, which is part of its total cross section. The physical dijet final state, described in Fig. 17, contains two jet cones of half angle δ and isotropic soft gluons within the energy resolution ϵQ, Q being the e^-e^+ invariant mass. The Born cross section is the same as the total one in Eq. (11). With the constrained phase space for real gluons, the infrared

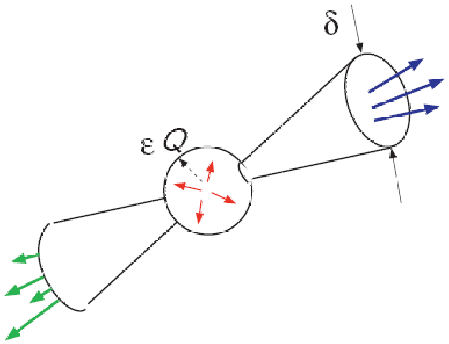


Fig. 17: Dijet final state in e^-e^+ annihilation.

cancellation is not complete, and logarithmic enhancement appears. The explicit NLO calculations imply that the isotropic soft gluons give a contribution proportional to $2\ln^2(2\epsilon Q/\mu)-\pi^2/6$, the collinear gluons in the cones with energy higher than the resolution give $-3\ln(Q\delta/\mu)-2\ln^2(2\epsilon)-4\ln(Q\delta/\mu)\ln(2\epsilon)+17/4-\pi^2/3$, and the virtual corrections contribute $-2\ln^2(Q/\mu)+3\ln(Q/\mu)-7/4+\pi^2/6$. The total NLO corrections indicate that the dijet cross section is infrared finite, but logarithmically enhanced:

$$3\ln\delta+4\ln\delta\ln(2\epsilon)+\frac{\pi^2}{3}-\frac{5}{2}, \tag{73}$$

where the double logarithm $\ln\delta\ln(2\epsilon)$ is attributed to the overlap of the collinear and soft logarithms.

4.1 Jet in Experiments

To describe the kinematics for jets, we define the pseudorapidity $\eta=\ln[\cot(\theta/2)]$, which is related to the polar angle θ with respect to the beam direction, and the azimuthal angle ϕ. That is, $\theta=0$, 90°, and 180° correspond to $\eta=+\infty$, 0 and $-\infty$, respectively. Comparison of theoretical and experimental descriptions for jet observables is nontrivial. One needs jet algorithms that map experimental measurements with theoretical calculations as close as possible. The infrared safety [61] is an important guideline for setting up a jet algorithm. There are two major classes of jet algorithms in the literature: cone algorithms and sequential algorithms. The former is a geometrical method, which stamps out jets on the η-ϕ plane as with a cookie cutter. The latter combines particle four-momenta one by one following given kinematic criteria.

I take the seeded cone algorithm as an example to explain the operation in the first class of jet algorithms, which aims at finding stable cones via an iterative procedure. Start from a seed particle i, and consider a set of particles j with separations smaller than jet cone of radius R,

$$\Delta R_{ij}^2\equiv(\eta_i-\eta_j)^2+(\phi_i-\phi_j)^2<R^2. \tag{74}$$

Calculate the new cone center J by summing all particle four-momenta in the cone. A stable cone is composed of a set of particles i satisfying $\Delta R_{iJ}<R$. If the cone is stable, the procedure stops. Otherwise, take J as a new seed, and repeat the above procedure.

However, the seeded cone algorithm suffers the problem of infrared divergences. Such a geometrical algorithm does not differentiate infrared gluons from energetic gluons, so final outcomes depend on soft radiation and collinear splitting. This problem can be illustrated by considering a system of two particles 1 and 2, separated by R_{12} with $R<R_{12}<2R$. Each of particles 1 and 2, taken as a seed, forms a stable jet. One then adds a soft gluon to this system. It is obvious that a virtual soft gluon exchanged between jets 1 and 2 does not change the outcome; namely, a virtual soft gluon contributes to the dijet cross section. On the contrary, adding a real soft seed between jets 1 and 2 will merge the two jets because of $R<R_{12}<2R$. Therefore, a real soft gluon contributes to the single jet cross section. As

a result, the soft divergences do not cancel between the virtual and real corrections. One may speculate that starting from the hardest particle may avoid the difficulty caused by the soft seed. It turns out that the collinear splitting would change the outcome. Including a more energetic particle into the above system, which is emitted between particles 1 and 2. Taking this central particle as the seed, one constructs a single stable jet formed by the three particles. A self-energy correction to the central particle does not change this final state, and contributes to the single jet cross section. However, the splitting of the central particle may produce two particles, which are less energetic than particles 1 and 2. Then one has to take particle 1 or 2 as the seed, and ends up with two stable jets. That is, the collinear splitting contributes to the dijet cross section, and there is no cancellation between virtual and real corrections. It is concluded that a seeded cone algorithm is not infrared safe.

Next I introduce sequential algorithms by taking the k_T algorithm as an example. For any pair of particles i and j, find the minimum of the following three distances

$$d_{ij} = \min(k_{Ti}^2, k_{Tj}^2)\frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = k_{Ti}^2, \quad d_{jB} = k_{Tj}^2, \tag{75}$$

with k_T being is a jet transverse momentum. If the minimum is d_{iB} or d_{jB}, i or j is a jet, and removed from the list of particles. Otherwise, i and j are merged into a new jet. Repeat the above procedure until no particles are left. The other sequential algorithms include the Cambridge/Aachen and anti-k_T ones with the definitions of the distances

$$\begin{aligned} d_{ij} &= \frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = 1, \quad d_{jB} = 1, \\ d_{ij} &= \min(k_{Ti}^{-2}, k_{Tj}^{-2})\frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = k_{Ti}^{-2}, \quad d_{jB} = k_{Tj}^{-2}, \end{aligned} \tag{76}$$

respectively. The grouping starts from soft (energetic) particles and usually leads to an irregular (round) jet shape in the k_T (anti-k_T) algorithm. Note that a sequential algorithm differentiates infrared gluons from energetic ones: adding a soft real gluon does not modify a cone center, so it does not change the outcome.

4.2 Jets in Theory

As outlined in the Introduction, we intend to establish a theoretical framework for jet study, following the idea of the factorization theorem for the DIS in Sec. 2. At NLO, a jet is produced in DIS, as the gluon emitted by the initial-state or final-state quark is collimated to the final-state quark. The restricted phase space of the final-state quark and the gluon in a small angular separation renders an incomplete cancellation between the virtual and real corrections. Hence, jet production is expected to be enhanced by collinear dynamics. Similarly, the initial-state quark propagator can be eikonalized in this collinear region, such that collinear gluons are detached from the initial-state quark and absorbed into a jet function. To all orders, the collinear gluons are collected by the Wilson link with the path-ordered exponential

$$W = \mathcal{P}\exp\left[-ig\int_0^\infty dzn\cdot A(zn)\right], \tag{77}$$

with an arbitrary vector n. The collinear gluon emitted by the final-state quark can be factorized into the jet function straightforwardly by applying the Fierz transformation. A more sophisticated factorization formula for the jet production in the DIS is then written as a convolution of a hard kernel H with a PDF and a jet function J. H denotes the contribution with the collinear pieces for the initial and final states being subtracted. This factorization formalism is the basis for the application of pQCD to jet physics.

The light-quark and gluon jet functions are defined by [62]

$$J_q(M_J^2, P_T, \nu^2, R, \mu^2) = \frac{(2\pi)^3}{2\sqrt{2}(P_J^0)^2 N_c}\sum_{N_J} Tr\left\{\not{n}\langle 0|q(0)W^{(\bar{q})\dagger}|N_J\rangle\langle N_J|W^{(\bar{q})}\bar{q}(0)|0\rangle\right\}$$

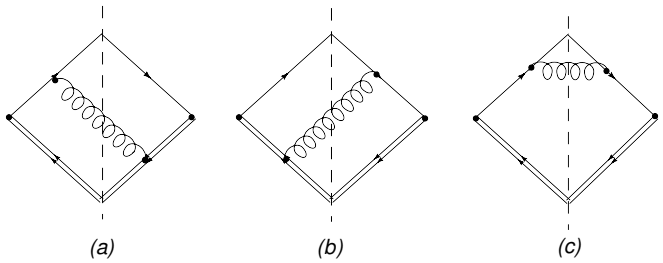


Fig. 18: Some NLO real corrections to the quark jet function.

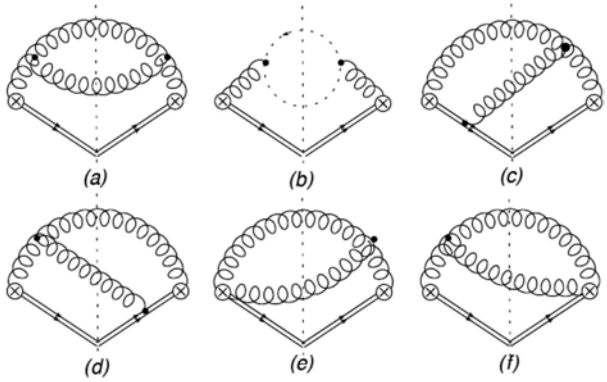


Fig. 19: Some NLO real corrections to the gluon jet function, where the dashed line represents a ghost field.

$$
\begin{aligned}
J_g(M_J^2, P_T, \nu^2, R, \mu^2) &= \times\delta(M_J^2 - \hat{M}_J^2(N_J, R))\delta^{(2)}(\hat{e} - \hat{e}(N_J))\delta(P_J^0 - \omega(N_J)), \\
&\quad \frac{(2\pi)^3}{2(P_J^0)^3 N_c} \sum_{N_J} \langle 0|\xi_\sigma F^{\sigma\nu}(0) W^{(g)\dagger}|N_J\rangle\langle N_J|W^{(g)} F_\nu^\rho(0)\xi_\rho|0\rangle \\
&\quad \times\delta(M_J^2 - \hat{M}_J^2(N_J, R))\delta^{(2)}(\hat{e} - \hat{e}(N_J))\delta(P_J^0 - \omega(N_J)), \qquad (78)
\end{aligned}
$$

where $|N_J\rangle$ denotes the final state with N_J particles within the cone of size R centered in the direction of the unit vector $\hat{e}$, $\hat{M}_J(N_J, R)$ ($\omega(N_J)$) is the invariant mass (total energy) of all N_J particles, and μ is the factorization scale. The above jet functions absorb the collinear divergences from all-order radiations associated with the energetic light jet of momentum $P_J^\mu = P_J^0 v^\mu$, in which P_J^0 is the jet energy, and the vector v is given by $v^\mu = (1, \beta, 0, 0)$ with $\beta = \sqrt{1 - (M_J/P_J^0)^2}$. $\xi^\mu = (1, -1, 0, 0)$ is a vector on the light cone. The coefficients in Eq. (78) have been chosen, such that the LO jet functions are equal to $\delta(M_J^2)$ in a perturbative expansion.

Underlying events include everything but hard scattering, such as initial-state radiation, final-state radiation, and multiple parton interaction (MPI). The Wilson lines in Eq. (78) have collected gluons radiated from both initial states and other final states in a scattering process, and collimated to the light-particle jets. Gluon exchanges between the quark fields q (or the gluon fields $F^{\sigma\nu}$ and F_ν^ρ) correspond to the final-state radiations. Both the initial-state and final-state radiations are leading-power effects in the factorization theorem, and have been included in the jet function definition. A chance of involving more partons in hard scattering is low, so the contribution from MPI is regarded as being subleading-power. This contribution should be excluded from data, but it is certainly difficult to achieve in experiments. Nevertheless, it still makes sense to compare predictions for jet observables based on Eq. (78) at the current leading-power accuracy with experimental data. At last, pile-up events must be removed in experiments [63], since they cannot be handled theoretically so far.

The NLO diagrams for the light-quark and gluon jet functions are displayed in Figs. 18 and 19,

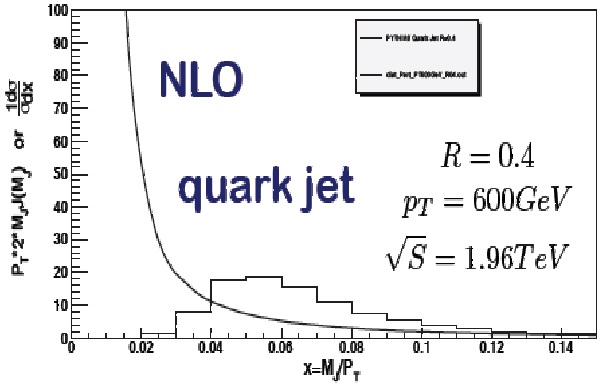


Fig. 20: Jet mass distribution at NLO.

respectively. Evaluating the jet functions up to NLO, a divergence is observed at small jet invariant mass M_J as shown in Fig. 20, that implies the nonperturbtive nature of the jet functions. The total NLO corrections in Mellin space indicate the existence of double logarithms, which hint the implementation of the resummation technique. Both the angular and energy resolutions are related to the jet mass: when M_J is not zero, particles in a jet cannot be completely collimated, and the jet must have finite minimal energy. This accounts for the source of the double logarithms. Recall that low p_T spectra of direct photons, dominated by soft and collinear radiations, are treated by the k_T resummation. The jet invariant mass is attributed to soft and collinear radiations, so the mass distribution can also be derived in the resummation formalism.

Varying the Wilson line direction n, we derive the differential equation for the light-quark jet function [64]

$$-\frac{n^2}{v\cdot n}v_\alpha\frac{d}{dn_\alpha}J_q(M_J^2,P_T,\nu^2,R,\mu^2)=2(K+G)\otimes J_q(M_J^2,P_T,\nu^2,R,\mu^2). \tag{79}$$

The above equation implies that the soft gluons in K are associated with the jet function J, a feature consistent with the anti-k_T algorithm. The solution to Eq. (79) resums the double logarithms in the jet function. One then convolutes the light-quark and gluon jet functions with the constituent cross sections of LO partonic dijet processes at the Tevatron and the PDF CTEQ6L [65]. The resummation predictions for the jet mass distributions at $R = 0.4$ and $R = 0.7$ are compared to the Tevatron CDF data [66] in Fig. 21 [67] with the kinematic cuts $P_T > 400$ GeV and the rapidity interval $0.1 < |Y| < 0.7$. The abbreviation NLL refers to the accuracy of the resummation, and NLO to the accuracy of the initial condition of the jet function solved from Eq. (79). The consistency of the resummation results with the CDF data is satisfactory.

4.3 Jet Substructure

It is known that a top quark produced almost at rest at the Tevatron can be identified by measuring isolated jets from its decay. However, this strategy does not work for identifying a highly-boosted top quark produced at the LHC. It has been observed that an ordinary high-energy QCD jet [68, 69] can have an invariant mass close to the top quark mass. A highly-boosted top quark, producing only a single jet, is then difficult to be distinguished from a QCD jet. This difficulty also appears in the identification of a highly-boosted new-physics resonance decaying into standard-model particles, or Higgs boson decaying into a bottom-quark pair. Hence, additional information needs to be extracted from jet internal structures in order to improve the jet identification at the LHC. The quantity, called planar flow [70], has been proposed for this purpose, which utilizes the geometrical shape of a jet: a QCD jet with large invariant mass mainly involves one-to-two splitting, so it leaves a linear energy deposition in a detector. A top-quark jet,

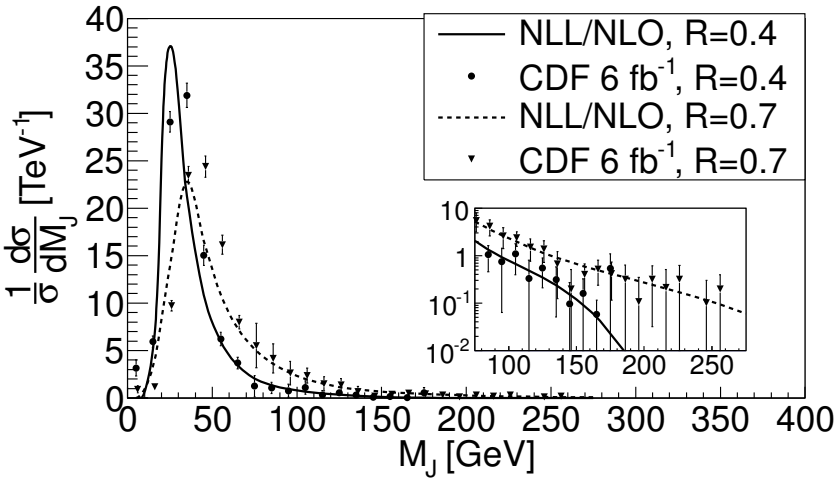


Fig. 21: Comparison of resummation predictions for the jet mass distributions to Tevatron CDF data with the kinematic cuts $P_T > 400$ GeV and $0.1 < |Y| < 0.7$ at $R = 0.4$ and $R = 0.7$. The inset shows the detailed comparison in large jet mass region.

proceeding with a weak decay, mainly involves one-to-three splitting, so it leaves a planar energy deposition. Measuring this additional information, it has been shown with event generators that the top-quark identification can be improved to some extent. Investigations on various observables associated with jet substructures are usually done using event generators. For a review on recent theoretical progress and the latest experimental results in jet substructures, see [71].

Here I focus on a jet substructure, called the energy profile, and explain how to calculate it in the resummation formalism [64]. This quantity describes the energy fraction accumulated in the cone of size r within a jet cone R, i.e., $r < R$. Its explicit definition is given by [72]

$$\Psi(r) = \frac{1}{N_J} \sum_J \frac{\sum_{r_i<r, i\in J} P_{Ti}}{\sum_{r_i<R, i\in J} P_{Ti}}, \tag{80}$$

with the normalization $\Psi(R) = 1$, where P_{Ti} is the transverse momentum carried by particle i in the jet J, and $r_i < r$ ($r_i < R$) means the flow of particle i into the jet cone r (R). Different types of jets are expected to exhibit different energy profiles. For example, a light-quark jet is narrower than a gluon jet; that is, energy is accumulated faster with r in a light-quark jet than in a gluon jet. A heavy-particle jet certainly has a distinct energy profile, which can be used for its identification. The importance of higher-order corrections and their resummation for studying a jet energy profile have been first emphasized in [73]. Another approach based on the soft-collinear effective theory and its application to jet production at an electron-positron collider can be found in Refs. [74–76].

We first define the jet energy functions $J_f^E(M_J^2, P_T, \nu^2, R, r)$ with $f = q(g)$ denoting the light-quark (gluon), which describe the energy accumulation within the cone of size $r < R$. The definition is chosen, such that $J_f^{E(0)} = P_T \delta(M_J^2)$ at LO. The Feynman rules for J_f^E are similar to those for the jet functions J_f at each order of α_s, except that a sum of the step functions $\sum_i k_i^0 \Theta(r - \theta_i)$ is inserted, where k_i^0 (θ_i) is the energy (the angle with respect to the jet axis) of particle i. For example, the jet energy functions J_f^E are expressed, at NLO, as

$$\begin{aligned} J_q^{E(1)}(M_J^2, P_T, \nu^2, R, r, \mu^2) &= \frac{(2\pi)^3}{2\sqrt{2}(P_J^0)^2 N_c} \sum_{\sigma,\lambda} \int \frac{d^3p}{(2\pi)^3 2p^0} \frac{d^3k}{(2\pi)^3 2k^0} \\ &\times [p^0 \Theta(r - \theta_p) + k^0 \Theta(r - \theta_k)] \\ &\times \mathrm{Tr}\left\{ \not{\xi} \langle 0|q(0) W^{(\bar{q})\dagger}|p, \sigma; k, \lambda\rangle \langle k, \lambda; p, \sigma| W^{(\bar{q})} \bar{q}(0)|0\rangle \right\} \end{aligned}$$

$$
\begin{aligned}
J_g^{E(1)}(M_J^2, P_T, \nu^2, R, r, \mu^2) &= \times\delta(M_J^2-(p+k)^2)\delta^{(2)}(\hat{e}-\hat{e}_{\mathbf{p+k}})\delta(P_J^0-p^0-k^0), \\
&\quad \frac{(2\pi)^3}{2(P_J^0)^3 N_c}\sum_{\sigma,\lambda}\int\frac{d^3p}{(2\pi)^3 2p^0}\frac{d^3k}{(2\pi)^3 2k^0} \\
&\quad \times[p^0\Theta(r-\theta_p)+k^0\Theta(r-\theta_k)] \\
&\quad \times\langle 0|\xi_\sigma F^{\sigma\nu}(0)W^{(g)\dagger}|p,\sigma;k,\lambda\rangle\langle k,\lambda;p,\sigma|W^{(g)}F_\nu^\rho(0)\xi_\rho|0\rangle \\
&\quad \times\delta(M_J^2-(p+k)^2)\delta^{(2)}(\hat{e}-\hat{e}_{\mathbf{p+k}})\delta(P_J^0-p^0-k^0), \qquad (81)
\end{aligned}
$$

where the expansion of the Wilson links in α_s is understood. The factorization scale is set to $\mu = P_T$ to remove the associated logarithms, so its dependence will be suppressed below.

The Mellin-transformed jet energy function $\bar{J}_q^E$ obeys a similar differential equation [64]

$$
-\frac{n^2}{v\cdot n}v_\alpha\frac{d}{dn_\alpha}\bar{J}_q^E(N=1,P_T,\nu^2,R,r) = 2(\bar{K}+G)\bar{J}_q^E(N=1,P_T,\nu^2,R,r), \qquad (82)
$$

which can be solved simply. Inserting the solutions to Eq. (82) into Eq. (80), the jet energy profile is derived. Note that a jet energy profile with $N = 1$ is not sensitive to the nonperturbative contribution, so the predictions are free of the nonperturbative parameter dependence, in contrast to the case of the jet invariant mass distribution. It has been found that the light-quark jet has a narrower energy profile than the gluon jet, as exhibited in Fig. 22 for $\sqrt{s} = 7$ TeV and the interval 80 GeV $< P_T <$ 100 GeV of the jet transverse momentum. The broader distribution of the gluon jet results from stronger radiations caused by the larger color factor $C_A = 3$, compared to $C_F = 4/3$ for a light-quark jet.

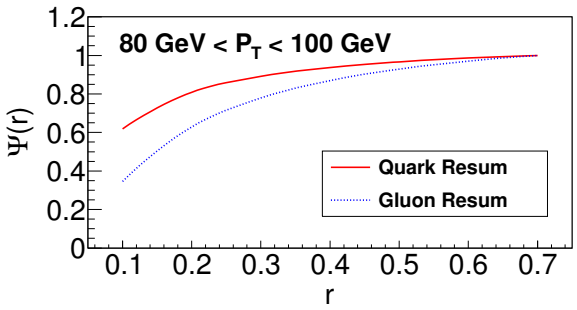


Fig. 22: Resummation predictions for the energy profiles of the light-quark (solid curve) and gluon (dotted curve) jets with $\sqrt{s} = 7$ TeV and 80 GeV $< P_T <$ 100 GeV.

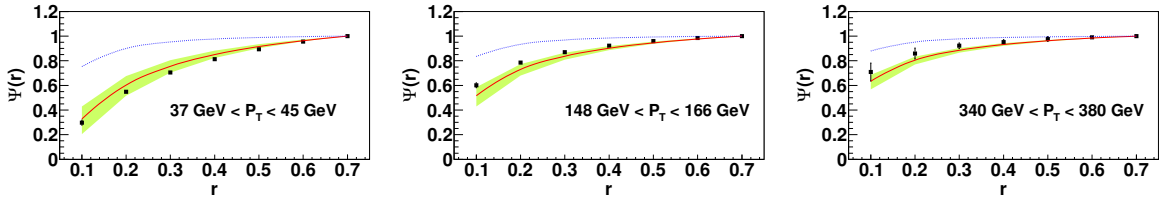


Fig. 23: Comparison of resummation predictions for the jet energy profiles with $R = 0.7$ to Tevatron CDF data in various P_T intervals. The NLO predictions denoted by the dotted curves are also displayed.

One then convolutes the light-quark and gluon jet energy functions with the constituent cross sections of the LO partonic subprocess and CTEQ6L PDFs [65] at certain collider energy. The predictions are directly compared with the Tevatron CDF data [72] as shown in Fig. 23. It is evident that the resummation predictions agree well with the data in all P_T intervals. The NLO predictions derived from $\bar{J}_f^{E(1)}(1, P_T, \nu_{\rm fi}^2, R, r)$ are also displayed for comparison, which obviously overshoot the data. The resummation predictions for the jet energy profiles are compared with the LHC CMS data at 7 TeV [77]

from the anti-k_T jet algorithm [78] in Fig. 24, which are also consistent with the data in various P_T intervals. Since one can separate the contributions from the light-quark jet and the gluon jet, the comparison with the CDF and CMS data implies that high-energy (low-energy) jets are mainly composed of the light-quark (gluon) jets. Hence, a precise measurement of the jet energy profile as a function of jet transverse momentum can be used to experimentally discriminate the production mechanism of jets in association with other particles, such as electroweak gauge bosons, top quarks and Higgs bosons.

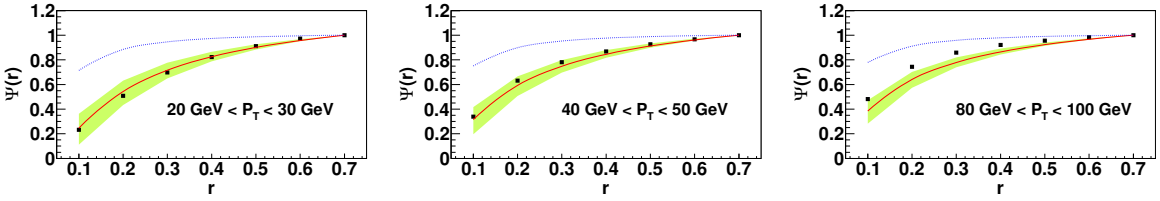


Fig. 24: Resummation predictions for the jet energy profiles with $R = 0.7$ compared to LHC CMS data in various P_T intervals. The NLO predictions denoted by the dotted curves are also displayed.

5 Hadronic heavy-quark decays

Hadronic decays of heavy-quark bound states, such as B, B_s, and Λ_b, are one of the focuses of LHCb physics, whose precision measurement may reveal new physics in the flavor sector. They are difficult to analyze theoretically because of complicated QCD dynamics and multiple characteristic scales they involve: the W boson mass m_W, the b quark mass m_b, and the QCD scale $\Lambda_{\rm QCD}$. The standard procedure is first to integrate out the scale m_W, such that QCD dynamics is organized into an effective weak Hamiltonian [79]. For the $B \to D\pi$ decays, the effective Hamiltonian is written as

$$\mathcal{H}_{\rm eff} = \frac{G_F}{\sqrt{2}} V_{cb}V_{ud}^* \Big[C_1(\mu)O_1(\mu) + C_2(\mu)O_2(\mu)\Big], \tag{83}$$

where G_F is the Fermi coupling constant, $V_{cb}V_{ud}^*$ is the product of the Cabibbo-Kobayashi-Maskawa matrix elements, μ is the renormalization scale, $C_{1,2}$ are the Wilson coefficients, and the four-fermion operators are defined by

$$O_1 = (\bar{d}b)_{V-A}(\bar{c}u)_{V-A}\,, \qquad O_2 = (\bar{c}b)_{V-A}(\bar{d}u)_{V-A}. \tag{84}$$

For exclusive processes, such as hadron form factors, the collinear factorization was developed in [80–83]. The range of a parton momentum fraction x, contrary to that in the inclusive case, is not experimentally controllable, and must be integrated over between 0 and 1. Hence, the end-point region with a small x is not avoidable. If there is no end-point singularity developed in a hard kernel, the collinear factorization works. If such a singularity occurs, indicating the breakdown of the collinear factorization, the k_T factorization should be employed, because the parton transverse momentum k_T is not negligible. To derive $B \to D\pi$ decay amplitudes, one evaluates the hadronic matrix elements $\langle D\pi|O_i(\mu)|B\rangle$. Different theoretical approaches have been developed for this purpose, which include the factorization assumption, the QCD-improved factorization, the perturbative QCD, the soft-collinear effective theory, the light-cone QCD sum rules, and the quark-diagram parametrization. In this section I briefly introduce the basic ideas of the first three approaches [24].

5.1 Factorization Assumption

Intuitively, decay products from a heavy b quark move fast without further interaction between them. This naive picture is supported by the color-transparency argument [84]: the Lorentz contraction renders energetic final states emitted from the weak vertex have small longitudinal color dipoles, which cannot

be resolved by soft gluons. Therefore, the hadronic matrix element $\langle O(\mu)\rangle$ is factorized into a product of two matrix elements of single currents, governed by decay constants and form factors, without soft gluon exchanges between them. This factorization assumption (FA) [14] was first proved in the framework of large energy effective theory [85], and justified in the large N_c limit [86]. For the $B \to D\pi$ decays, the color-allowed (color-suppressed) amplitude, involving the $B \to D$ ($B \to \pi$) transition form factor, is proportional to the Wilson coefficient $a_1 = C_2 + C_1/N_c$ ($a_2 = C_1 + C_2/N_c$).

In spite of its simplicity, the FA encounters three principal difficulties. First, a hadronic matrix element under the FA is independent of the renormalization scale μ, as the vector or axial-vector current is partially conserved. Consequently, the amplitude $C(\mu)\langle O\rangle_{\rm fact}$ is not truly physical as the scale dependence of the Wilson coefficient does not get compensation from the matrix element. This problem may not be serious for color-allowed modes, because the parameter a_1 is roughly independent of μ. It is then not a surprise that the simple FA gives predictions in relatively good agreement with data of these modes. However, the parameter a_2 depends strongly on the renormalization scale and on the renormalization scheme, because of the similar magnitude and different sign of the $C_1(\mu)$ and $C_2(\mu)/N_c$ terms (calculated in the NDR scheme and for $\Lambda^{(5)}_{\overline{MS}} = 225$ GeV, the Wilson coefficients have the values $C_1(m_B) = -0.185$ and $C_2(m_B) = 1.082$ [79], m_B being the B meson mass). This may be the reason why the FA fails to accommodate data of color-suppressed modes. It also means that a_2 is more sensitive to subleading contributions.

The second difficulty is related to the first one: nonfactorizable effects have been neglected in the FA. This neglect may be justified for color-allowed modes due to the large and roughly μ-independent value of a_1, but not for color-suppressed modes, such as $B \to J/\psi K^{(*)}$. The J/ψ meson emitted from the weak vertex is not energetic, and the color-transparency argument does not apply. To circumvent this difficulty, nonfactorizable contributions were parameterized into the parameters χ_i [87, 88],

$$\begin{aligned} a_1^{\rm eff} &= C_2(\mu) + C_1(\mu)\left[\frac{1}{N_c} + \chi_1(\mu)\right], \\ a_2^{\rm eff} &= C_1(\mu) + C_2(\mu)\left[\frac{1}{N_c} + \chi_2(\mu)\right]. \end{aligned} \tag{85}$$

The μ dependence of the Wilson coefficients is assumed to be exactly compensated by that of $\chi_i(\mu)$ [89]. It is obvious that the introduction of χ_i does not really resolve the scale problem in the FA.

Third, strong phases are essential for predicting CP asymmetries in exclusive B meson decays. These phases, arising from the Bander-Silverman-Soni (BSS) mechanism [90], are ambiguous in the FA: the charm quark loop contributes an imaginary piece proportional to

$$\int du u(1-u)\theta(q^2u(1-u) - m_c^2), \tag{86}$$

where q^2 is the invariant mass of the gluon attaching to the charm loop. Since q^2 is not precisely defined in the FA, one cannot obtain definite information of strong phases from Eq. (86). Moreover, it is legitimate to question whether the BSS mechanism is an important source of strong phases in B meson decays. Viewing the above difficulties, the FA is not a complete model, and it is necessary to go beyond the FA by developing reliable and systematic theoretical approaches.

5.2 QCD-improved Factorization

The color-transparency argument allows the addition of hard gluons between the energetic mesons emitted from the weak vertex and the B meson transition form factors. These hard gluon exchanges lead to higher-order corrections in the coupling constant α_s to the FA. By means of Feynman diagrams, they appear as the vertex corrections in the first two rows of Fig. 25 [15]. It has been shown that soft divergences cancel among them, when computed in the collinear factorization theorem. These $O(\alpha_s)$

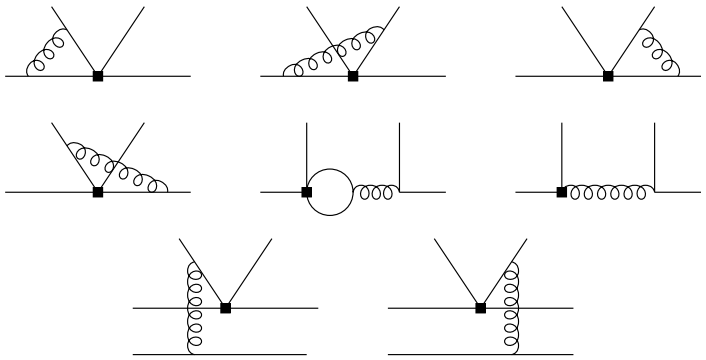


Fig. 25: $O(\alpha_s)$ corrections to the FA in the QCDF approach.

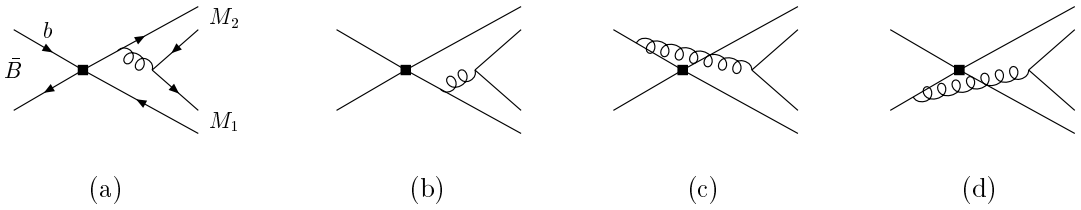


Fig. 26: Annihilation contributions.

corrections weaken the μ dependence in the Wilson coefficients, and generate strong phases. Besides, hard gluons can also be added to form the spectator diagrams in the last row of Fig. 25. Feynman rules of these two diagrams differ by a minus sign in the soft region resulting from the involved quark and anti-quark propagators. Including the above nonfactorizable corrections to the FA leads to the QCD-improved factorization (QCDF) approach [15]. The gluon invariant mass q^2 in the BSS mechanism can be unambiguously defined and related to parton momentum fractions in QCDF. Hence, the theoretical difficulties in the FA are resolved. This is a breakthrough towards a rigorous framework for two-body hadronic B meson decays in the heavy quark limit.

Corrections in higher powers of $1/m_b$ to the FA can also be included into QCDF, such as those from the annihilation topology in Fig. 26, and from twist-3 contributions to the spectator amplitudes. However, it has been found that endpoint singularities exist in these high-power contributions, which arise from the divergent integral $\int_0^1 dx/x$, x being a momentum fraction. These singularities have the same origin as those in the collinear collinear factorization formulas for B meson transition form factors [91]. Because of the endpoint singularities, the annihilation and twist-3 spectator contributions must be parameterized as [15]

$$\ln\frac{m_B}{\Lambda_h}\left(1+\rho_A e^{i\delta_A}\right)\ , \quad \ln\frac{m_B}{\Lambda_h}\left(1+\rho_H e^{i\delta_H}\right)\ , \tag{87}$$

respectively, with the hadronic scale Λ_h. A QCDF formula then contains the arbitrary parameters $\rho_{A,H}$ and $\delta_{A,H}$. Setting these parameters to zero, one obtains predictions in the "default" scenario, and the variation of the arbitrary parameters gives theoretical uncertainties. If tuning these parameters to fit data, one obtains results in the scenarios "S", "S2",... [92].

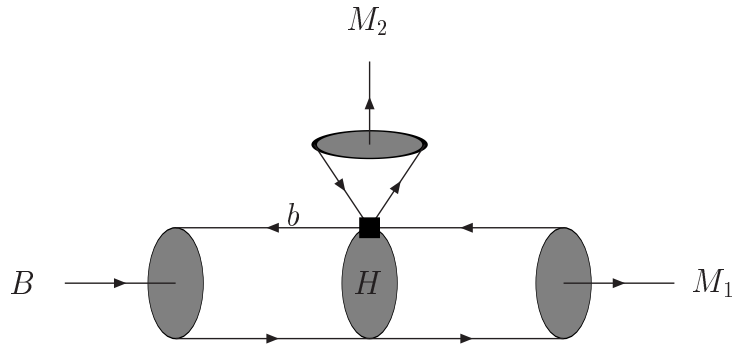


Fig. 27: Perturbative QCD factorization.

5.3 Perturbative QCD

The endpoint singularities signal the breakdown of the collinear factorization for two-body hadronic B meson decays. Motivated by the removal of these singularities, the perturbative QCD (PQCD) approach based on the k_T factorization theorem was developed [16–19]. A parton transverse momentum k_T can be generated by gluon radiations, before hard scattering occurs. The endpoint singularities from the small x region simply indicate that k_T is not negligible. Taking into account k_T, a particle propagator does not diverge as $x \to 0$. The B meson transition form factors, and the spectator and annihilation contributions are then all calculable in the framework of the k_T factorization theorem. It has been shown that a $B \to M_1 M_2$ decay amplitude is factorized into the convolution of the six-quark hard kernel, the jet function and the Sudakov factor with the bound-state wave functions as shown in Fig. 27,

$$A(B \to M_1 M_2) = \phi_B \otimes H \otimes J \otimes S \otimes \phi_{M_1} \otimes \phi_{M_2}. \tag{88}$$

The jet function J comes from the threshold resummation, which exhibits suppression in the small x region [93]. The Sudakov factor S comes from the k_T resummation, which exhibits suppression in the small k_T region [39, 40]. These resummation effects guarantee the removal of the endpoint singularities. J (S), organizing double logarithms in the hard kernel (meson wave functions), is hidden in H (the three meson states) in Fig. 27. The arbitrary parameters introduced in QCDF are not necessary, and PQCD involves only universal and controllable inputs.

The theoretical difficulties in the FA are also resolved in PQCD but in a different manner. The FA limit of the PQCD approach at large m_b, which is not as obvious as in QCDF, has been examined [93]. It was found that the factorizable emission amplitude decreases like $m_b^{-3/2}$, if the B meson decay constant f_B scales like $f_B \propto m_b^{-1/2}$. This power-law behavior is consistent with that obtained in [15, 94]. The higher-order corrections to the FA have been included in PQCD, which moderate the dependence on the renormalization scale μ. The ratio of the spectator contribution over the factorizable emission contribution decreases with m_b in PQCD, showing a behavior close to that in QCDF. The gluon invariant mass q^2 in the BSS mechanism is unambiguously defined and related to parton momentum fractions. The penguin annihilation amplitude is almost imaginary in PQCD [18], whose mechanism is similar to the BSS one [90]: in the annihilation topology, the loop is formed by the internal particles in the LO hard kernel and by infinitely many Sudakov gluons exchanged between two partons in a light meson. A sizable strong phase is generated, when the internal particles go on mass shell. In terms of the principal-value prescription for the internal particle propagator, the strong phase is given by [18]

$$\frac{1}{x m_B^2 - k_T^2 + i\epsilon} = \frac{P}{x m_B^2 - k_T^2} - i\pi\delta(x m_B^2 - k_T^2). \tag{89}$$

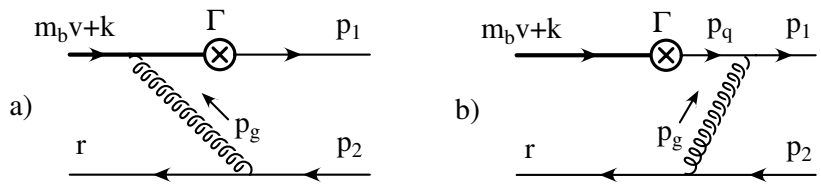


Fig. 28: Diagrams for the $B \to \pi$ form factor in QCD.

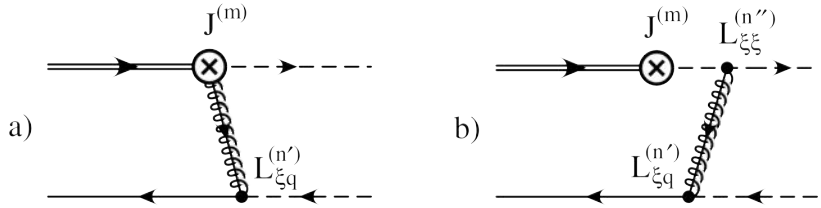


Fig. 29: Diagrams for the $B \to \pi$ form factor in $\mathrm{SCET_I}$.

5.4 Soft-Collinear Effective Theory

The soft-collinear effective theory (SCET) based on the collinear factorization is formulated in the framework of OPE [20–23]. The matching at different scales involved in B meson decays has been carefully handled in SCET. Take the simple $B \to \pi$ transition form factor in Fig. 28 as an example. The soft spectator in the B meson carries the momentum $r \sim O(\Lambda_{\rm QCD})$, because it is dominated by soft dynamics. If the spectator in the energetic pion carries the momentum $p_2 \sim O(m_b)$, the virtual gluon in Fig. 28 is off-shell by $p_g^2 = (p_2 - r)^2 = -2p_2 \cdot r \sim O(m_b \Lambda_{\rm QCD})$. Then the virtual quark in Figs. 28(a) is off-shell by $(m_b v + k + p_g)^2 - m_b^2 \sim O(m_b^2)$, where v is the b quark velocity and $k \sim O(\Lambda_{\rm QCD})$ denotes the Fermi motion of the b quark. Hence, B meson decays contain three scales below m_W: m_b, $\sqrt{m_b \Lambda_{\rm QCD}}$, and $\Lambda_{\rm QCD}$.

The separate matching at the two scales m_b and $\sqrt{m_b \Lambda_{\rm QCD}}$ is briefly explained below [95]. The first step is to integrate out the lines off-shell by m_b^2 in QCD, and the resultant effective theory is called $\mathrm{SCET_I}$. One then derives the zeroth-order effective current $J^{(0)}$ from the $b \to u$ weak vertex, and the first-order effective current $J^{(1)}$ by shrinking the virtual b quark line in Fig. 28(a). The next step is to integrate out the lines off-shell by $m_b \Lambda_{\rm QCD}$ in $\mathrm{SCET_I}$, arriving at $\mathrm{SCET_{II}}$. The relevant diagrams to start with are displayed in Fig. 29. Shrinking all the lines off-shell by $m_b \Lambda_{\rm QCD}$, one derives the corresponding Wilson coefficients, i.e., the jet functions, and the effective four-fermion operators. Sandwiching these four-fermion operators by the initial B meson state and the final pion state leads to the B meson and pion distribution amplitudes. The $B \to \pi$ transition form factor is then factorized as depicted in Fig. 30. The factorization of two-body hadronic B meson decays is constructed in a similar way, and the result is also shown in Fig. 30.

At leading power in $1/m_b$, there is no large source of strong phases in SCET (the annihilation contribution is parametrically power-suppressed). To acquire strong phases, it has been argued that $c\bar{c}$ (charming) penguins could give long-distance effects at leading power [96]. This contribution is nonperturbative, so it must be parameterized as an arbitrary amplitude $A^{c\bar{c}}$. Including the charming penguin, SCET has been applied as an QCD-improved parametrization, and $A^{c\bar{c}}$ is determined together with other hadronic inputs from data. It should be mentioned that the long-distance charming-penguin contribution is power-suppressed according to QCDF, PQCD and light-cone sum rules [97].

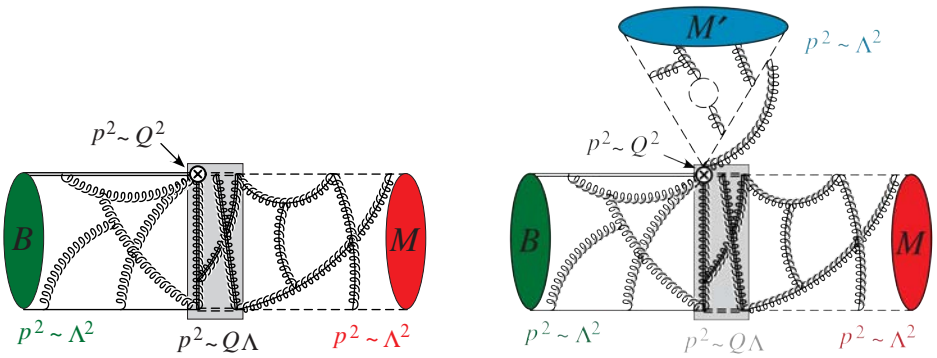


Fig. 30: Factorization of the $B \to \pi$ form factor and of the $B \to M_1 M_2$ decay in SCET.

5.5 Puzzles in B Physics

Before concluding, I review the long-standing puzzles in hadronic two-body B meson decays, which have not yet been fully resolved so far. According to a naive estimate of the color-suppressed tree amplitude, the hierarchy of the branching ratios $B(B^0 \to \pi^0\pi^0) \sim O(\lambda^2) B(B^0 \to \pi^\mp\pi^\pm)$ with the CKM parameter $\lambda \approx 0.2$ is expected. However, the data [98]

$$\begin{aligned} B(B^0 \to \pi^\mp\pi^\pm) &= (5.10 \pm 0.19) \times 10^{-6} , \\ B(B^0 \to \pi^0\pi^0) &= (1.91^{+0.22}_{-0.23}) \times 10^{-6} , \end{aligned} \tag{90}$$

imply $B(B^0 \to \pi^0\pi^0) \sim O(\lambda) B(B^0 \to \pi^\mp\pi^\pm)$, giving rise to the $B \to \pi\pi$ puzzle. As observed in [99], the NLO corrections, despite of increasing the color-suppressed tree amplitude significantly, are not enough to enhance the $B^0 \to \pi^0\pi^0$ branching ratio to the measured value. A much larger color-suppressed tree amplitude, about the same order as the color-allowed tree amplitude, must be obtained in order to resolve the puzzle [100,101]. To make sure that the above NLO effects are reasonable, the PQCD formalism has been applied to the $B \to \rho\rho$ decays [99], which also receive the color-suppressed tree contribution. It was observed that the NLO PQCD predictions are in agreement with the data $B(B^0 \to \rho^0\rho^0) = (0.73^{+0.27}_{-0.28}) \times 10^{-6}$ [98]. One concludes that it is unlikely to accommodate the measured $B^0 \to \pi^0\pi^0$ and $\rho^0\rho^0$ branching ratios simultaneously in PQCD, and that the $B \to \pi\pi$ puzzle remains.

It has been claimed that the $B \to \pi\pi$ puzzle has been resolved in the QCDF approach [15] with an input from SCET [102–104]: the inclusion of the NLO jet function, the hard coefficient of SCET_{II}, into the QCDF formula for the color-suppressed tree amplitude gives sufficient enhancement of the $B^0 \to \pi^0\pi^0$ branching ratio, if adopting the parameter scenario "S4" [105]. It is necessary to investigate whether the proposed new mechanism deteriorates the consistency of theoretical results with other data. The formalism in [102] has been extended to the $B \to \rho\rho$ decays as a check [99]. It was found that the NLO jet function overshoots the observed $B^0 \to \rho^0\rho^0$ branching ratio very much as adopting "S4". That is, it is also unlikely to accommodate the $B \to \pi\pi$ and $\rho\rho$ data simultaneously in QCDF.

Table 1: Polarization fractions in the penguin-dominated $B \to VV$ decays.

Mode	BABAR	Belle
ϕK^{*+}	$0.49 \pm 0.05 \pm 0.03$	$0.52 \pm 0.08 \pm 0.03$
$K^{*+}\rho^0$	$0.78 \pm 0.12 \pm 0.03$	
$K^{*0}\rho^+$	$0.52 \pm 0.10 \pm 0.04$	$0.43 \pm 0.11^{+0.05}_{-0.02}$
$K^{*+}K^{*0}$	$0.75^{+0.16}_{-0.26} \pm 0.03$	

For penguin-dominated $B \to VV$ decays, such as those listed in Table 1 [98], the polarization fractions deviate from the naive counting rules based on kinematics [106]. This is the so-called the $B \to \phi K^*$ puzzle. Many attempts to resolve the $B \to \phi K^*$ polarizations have been made [107], which include new physics [108–112], the annihilation contribution [113, 114] in the QCDF approach, the charming penguin in SCET [115], the rescattering effect [116–118], and the $b \to sg$ (the magnetic penguin) [119] and $b \to s\gamma$ [120] transitions. The annihilation contribution from the scalar penguin operators improves the consistency with the data, because it is of the same order for all the three final helicity states, and could enhance the transverse polarization fractions [106]. However, the PQCD analysis of the scalar penguin annihilation amplitudes indicates that the $B \to \phi K^*$ puzzle cannot be resolved completely [107]. A reduction of the $B \to K^*$ form factor A_0, which is associated with the longitudinal polarization, further helps accommodating the data [121].

The penguin-dominated $B \to K^*\rho$ decays are expected to exhibit similar polarization fractions. This is the reason why the longitudinal polarization fraction in the $B^+ \to K^{*0}\rho^+$ decay, which contains only the penguin contribution, is close to $f_L(\phi K^*) \sim 0.5$ as listed in Table 1. Another mode $B^+ \to K^{*+}\rho^0$, nevertheless, exhibits a large longitudinal polarization fraction around 0.8. This mode involves tree amplitudes, which are subdominant, and should not cause a significant deviation from $f_L \sim 0.5$. Though the data of $f_L(K^{*0}\rho^+)$ from BABAR still suffer a large error, the different longitudinal polarization fractions, $f_L(K^{*+}\rho^0) \neq f_L(K^{*0}\rho^+)$, call for a deeper understanding. The $B^+ \to K^{*+}K^{*0}$ decay shows a longitudinal polarization fraction smaller than unity, but larger than 0.5. A more thorough study of the $B \to K^*K^*$ decays can help discriminating the various resolutions for the $B \to \phi K^*$ puzzle [121, 122].

The $B^0 \to K^\pm\pi^\mp$ decays depend on the tree amplitude T and the QCD penguin amplitude P. The data of the direct CP asymmetry $A_{CP}(B^0 \to K^\pm\pi^\mp) \approx -10\%$ then imply a sizable relative strong phase between T and P, which verifies the LO PQCD prediction made years ago [18]: the scalar penguin annihilation provides an important source of strong phases. The PQCD predictions for significant penguin annihilation have been confirmed by the recent measurement of the pure annihilation mode, $B(B_s \to \pi^+\pi^-) = (0.73 \pm 0.14) \times 10^{-6}$, which is consistent with 0.57×10^{-6} obtained in the LO PQCD approach [123]. The $B^\pm \to K^\pm\pi^0$ decays contain the additional color-suppressed tree amplitude C and electroweak penguin amplitude P_{ew}. Since both C and P_{ew} are subdominant, the approximate equality for the direct CP asymmetries $A_{CP}(B^\pm \to K^\pm\pi^0) \approx A_{CP}(B^0 \to K^\pm\pi^\mp)$ is expected. However, this naive expectation is in conflict with the data [98],

$$\begin{aligned} A_{CP}(B^0 \to K^\pm\pi^\mp) &= -0.086 \pm 0.007 \\ A_{CP}(B^\pm \to K^\pm\pi^0) &= 0.040 \pm 0.021, \end{aligned} \tag{91}$$

making the $B \to K\pi$ puzzle.

While LO PQCD gives a negligible C [18, 19], it is possible that this supposedly tiny amplitude receives a significant subleading correction. Note that the small C is attributed to the accidental cancellation between the Wilson coefficients C_1 and C_2/N_c at the scale of m_b. In [124] the important NLO contributions to the $B \to K\pi$ decays from the vertex corrections, the quark loops, and the magnetic penguins were calculated. It was observed that the vertex corrections increase C by a factor of 3, and induce a large phase about -80^o relative to T. The large and imaginary C renders the total tree amplitude $T + C$ more or less parallel to the total penguin amplitude $P + P_{ew}$ in the $B^\pm \to K^\pm\pi^0$ decays, leading to nearly vanishing $A_{CP}(B^\pm \to K^\pm\pi^0) = (-1^{+3}_{-6})\%$ at NLO (it is about -8% at LO). One concludes that the $B \to K\pi$ puzzle has been alleviated, but not yet gone away completely. Whether new physics effects [125, 126] are needed will become clear when the data get precise. More detailed discussion on this subject can be found in [127].

6 Summary

Despite of nonperturbative nature of QCD, theoretical frameworks with predictive power can be developed. They are based on the factorization theorems, in which nonperturbative dynamics is absorbed into PDFs, and the remaining infrared finite contributions go to hard kernels. A PDF is universal (process-independent) and can be extracted from data, while a hard kernel is calculable in perturbation theory. Both the collinear and k_T factorization theorems are the fundamental tools of pQCD. The collinear factorization theorem is a simpler version, and has been intensively studied and widely applied. The k_T factorization theorem is more complicated, and many of its aspects have not been completely explored.

Sophisticated evolution equations and resummation techniques have been developed in pQCD, which enhance predictive power, and increase theoretical precision. All the known single- and double-logarithm summations, including their unifications, have been explained in the CSS resummation formalism. The point is the treatment of real gluon emissions under different kinematic orderings, and the resultant logarithmic summations are summarized in Table 2. The k_T and threshold resummations, and the DGLAP and BFKL equations have been applied to various QCD processes.

Table 2: Single- and double-logarithmic summations under different kinematic orderings.

	small x	intermediate x	large x
rapidity ordering	BFKL equation	k_T resummation	
k_T ordering		DGLAP equation	threshold resummation
angular ordering	CCFM	equation; joint	resummation

Experimental and theoretical studies of jet physics have been reviewed. Especially, it was pointed out that jet substructures could be calculated in pQCD: starting with the jet function definition, applying the factorization theorem and the resummation technique, one can predict observables, which are consistent with data. Because fixed-order calculations are not reliable at small jet invariant mass, and event generators have ambiguities, pQCD provides an alternative approach, that resolves the above difficulties. The pQCD formalism will improve the jet identification and new particle search at the LHC.

We have been able to go beyond the factorization assumption fr hadronic two-body heavy-quark decays by including QCD corrections. Different approaches have been discussed and commented: in QCDF the high-power corrections must be parameterized due to the existence of the endpoint singularities. There are no endpoint singularities in PQCD, which is based on the k_T factorization theorem, and in SCET, which employs the zero-bin subtraction [128]. A major difference arises from the treatment of the annihilation contribution, which is parameterized in QCDF and neglected in SCET, but is the main source of strong phases in PQCD.

Many subtle subjects on pQCD deserve more exploration, including the legitimate definition of TMDs, the gauge invariance of the k_T factorization, resummations of other types of logarithms, such as rapidity logarithms, non-global logarithms, and etc., jet substructures of boosted heavy particles, and the long-standing puzzles in B physics. pQCD remains as one of the most challenging research fields in high-energy physics.

Acknowledgment

This work was supported in part by the National Science Council of R.O.C. under Grant No. NSC-101-2112-M-001-006-MY3, and by the National Center for Theoretical Sciences of R.O.C.. The author acknowledges the hospitality of the organizers during the First Asia-Europe-Pacific School of High-energy Physics at Fukuoka, Japan in Oct., 2012.

References

[1] G. Sterman, arXiv:hep-ph/9606312.

[2] J.C. Collins and G. Sterman, Nucl. Phys. **B 185**, 172 (1981); J.C. Collins and D.E. Soper, Nucl. Phys. **B194**, 445 (1982).

[3] V.N. Gribov and L.N. Lipatov, Sov. J. Nucl. Phys. **15**, 438 (1972); G. Altarelli and G. Parisi, Nucl. Phys. **B126**, 298 (1977); Yu.L. Dokshitzer, Sov. Phys. JETP **46**, 641 (1977).

[4] E.A. Kuraev, L.N. Lipatov and V.S. Fadin, Sov. Phys. JETP **45**, 199 (1977); Ya.Ya. Balitsky and L.N. Lipatov, Sov. J. Nucl. Phys. **28**, 822 (1978); L.N. Lipatov, Sov. Phys. JETP **63**, 904 (1986).

[5] G. Sterman, Phys. Lett. B **179**, 281 (1986); Nucl. Phys. **B281**, 310 (1987).

[6] S. Catani and L. Trentadue, Nucl. Phys. **B327**, 323 (1989); Nucl. Phys. **B353**, 183 (1991).

[7] G.P. Korchemsky and G. Marchesini, Nucl. Phys. **B406**, 225 (1993); Phys. Lett. B **313**, 433 (1993).

[8] J.C. Collins and D.E. Soper, Nucl. Phys. **B193**, 381 (1981).

[9] J.C. Collins, D.E. Soper, and G. Sterman, Nucl. Phys. **B250**, 199 1985.

[10] M. Ciafaloni, Nucl. Phys. **B296**, 49 (1988); S. Catani, F. Fiorani, and G. Marchesini, Phys. Lett. B **234**, 339 (1990); Nucl. Phys. **B336**, 18 (1990); G. Marchesini, Nucl. Phys. **B445**, 49 (1995).

[11] H.-n. Li, Phys. Lett. B**454**, 328 (1999).

[12] E. Laenen, G. Sterman, and W. Vogelsang, Phys. Rev. Lett. **84**, 4296 (2000); Phys. Rev. D **63**, 114018 (2001).

[13] J. Shelton, arXiv:1302.0260 [hep-ph].

[14] M. Bauer, B. Stech, M. Wirbel, Z. Phys. C **29**, 637 (1985); *ibid.* **34**, 103 (1987).

[15] M. Beneke, G. Buchalla, M. Neubert, and C.T. Sachrajda, Phys. Rev. Lett. **83**, 1914 (1999); Nucl. Phys. **B591**, 313 (2000); Nucl. Phys. **B606**, 245 (2001).

[16] H-n. Li and H.L. Yu, Phys. Rev. Lett. **74**, 4388 (1995); Phys. Lett. B **353**, 301 (1995); Phys. Rev. D **53**, 2480 (1996).

[17] C.H. Chang and H-n. Li, Phys. Rev. D **55**, 5577 (1997).

[18] Y.Y. Keum, H-n. Li, and A.I. Sanda, Phys Lett. B **504**, 6 (2001); Phys. Rev. D **63**, 054008 (2001).

[19] C.D. Lü, K. Ukai, and M.Z. Yang, Phys. Rev. D **63**, 074009 (2001).

[20] C.W. Bauer, S. Fleming, and M. Luke, Phys. Rev. D **63**, 014006 (2001).

[21] C.W. Bauer, S. Fleming, D. Pirjol, and I.W. Stewart, Phys. Rev. D **63**, 114020 (2001).

[22] C.W. Bauer and I.W. Stewart, Phys. Lett. B **516**, 134 (2001).

[23] C.W. Bauer, D. Pirjol and I.W. Stewart, Phys. Rev. D **65**, 054022 (2002).

[24] H.-n. Li, Prog. Part. Nucl. Phys. **51**, 85 (2003).

[25] G. Sterman, *An Introduction to Quantum Field Theory*, Cambridge, 1993.

[26] T. Kinoshta, J. Math. Phys **3**, 650 (1962); T.D. Lee and M. Nauenberg, Phys. Rev. **133**, B1549 (1964).

[27] H.-L. Lai, et al., Phys. Rev. D **82**, 074024 (2010)..

[28] Nadolsky et al., arXiv:1206.3321 [hep-ph].

[29] S. Catani, M. Ciafaloni and F. Hautmann, Phys. Lett. B **242**, 97 (1990); Nucl. Phys. **B366**, 135 (1991).

[30] J.C. Collins and R.K. Ellis, Nucl. Phys. **B360**, 3 (1991).

[31] E.M. Levin, M.G. Ryskin, Yu.M. Shabelskii, and A.G. Shuvaev, Sov. J. Nucl. Phys. **53**, 657 (1991).

[32] X. Ji, and F. Yuan, Phys. Lett. B **543**, 66 (2002); A.V. Belitsky, X. Ji, and F. Yuan, Nucl. Phys. **B656**, 165 (2003).

[33] I.O. Cherednikov and N.G. Stefanis, Nucl. Phys. **B802**, 146 (2008).

[34] H.-n. Li, arXiv:hep-ph/9803202.

[35] J.C. Collins, Acta. Phys. Polon. B **34**, 3103 (2003).
[36] H.-n. Li and S. Mishima, Phys. Lett. B **674**, 182 (2009).
[37] G.P. Korchemsky and G. Sterman, Phys. Lett. B **340**, 96 (1994).
[38] H.-n. Li, Phys. Lett. B369, 137 (1996); Phys. Rev. D **55**, 105 (1997).
[39] J. Botts and G. Sterman, Nucl. Phys. **B325**, 62 (1989).
[40] H.-n. Li and G. Sterman, Nucl. Phys. **B381**, 129 (1992); H.-n. Li, Phys. Rev. D **48**, 4243 (1993).
[41] C. Corianó and H.-n. Li, Phys. Lett. B309, 409 (1993); Nucl. Phys. **B434**, 535 (1995).
[42] H.-n. Li and H.L. Yu, Phys. Rev. Lett. 74, 4388 (1995); Phys. Lett. B353, 301 (1995); H.-n. Li, Phys. Lett. B348, 597 (1995).
[43] H.-n. Li, Phys. Rev. D52, 3958 (1995); C.Y. Wu, T.W. Yeh and H.-n. Li, Phys. Rev. D **53**, 4982 (1996).
[44] J.C. Collins, Adv. Ser. Direct. High Energy Phys. **5**, 573 (1989).
[45] J. Kodaira and L. Trentadue, Phys. Lett. B **112**, 66 (1982).
[46] H.L. Lai and H.-n. Li, Phys. Rev. D **58**, 114020 (1998).
[47] H.L. Lai *et al.*, Phys. Rev. D **55**, 1280 (1997).
[48] T. Becher, C. Lorentzen, and M.D. Schwartz, Phys. Rev. D **86**, 054026 (2012).
[49] H-n. Li, Chin. J. Phys. **37**, 8 (1999); Phys. Lett. B **405**, 347 (1997).
[50] J. Kwieciński, A.D. Martin, and P.J. Sutton, Phys. Rev. D **53**, 6094 (1996).
[51] M. Froissart, Phys. Rev. **123**, 1053 (1961).
[52] V.S. Fadin and L.N. Lipatov, Phys. Lett. B **429**, 127 (1998); G. Camici and M. Ciafaloni, Phys. Lett. B **430**, 349 (1998).
[53] R.S. Thorne, Phys. Rev. D **60**, 054031 (1999).
[54] M. Beneke, P. Falgari, S. Klein, and C. Schwinn, Nucl. Phys. **B855**, 695 (2012).
[55] ATLAS Collaboration, Conference note atlas-conf-2011-121, 2011. https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2011-121/ATLAS-CO.
[56] R. Rajkhowa and J.K. Sarma, arXiv:1209.4350 [hep-ph].
[57] A.D. Martin, R.G. Roberts, W.J. Stirling, and R.S. Thorne, Eur. Phys. J. C **23**, 73 (2002).
[58] M. Devee, R. Baishya, and J.K. Sarma, Eur. Phys. J. C **72**, 2036 (2012).
[59] M. Arneodo et al., Nucl. Phys. **B483**, 3 (1997).
[60] G. Bozzi, B. Fuks, and M. Klasen, Nucl. Phys. **B794**, 46 (2008).
[61] G. Sterman and S. Weinberg, Phys. Rev. Lett. **39**, 1436 (1977).
[62] L.G. Almeida et al., Phys. Rev. D **79**, 074012 (2009).
[63] G. Soyez, G.P. Salam, J. Kim, S. Dutta, and M. Cacciari, arXiv:1211.2811 [hep-ph].
[64] H.-n. Li, Z. Li, Zhao, and C.-P. Yuan, Phys. Rev. Lett. **107**, 152001 (2011).
[65] J. Pumplin et al., JHEP **07**, 012 (2002).
[66] T. Aaltonen et al. (CDF Collaboration), Phys. Rev. D **85**, 091101 (2012).
[67] H.-n. Li, Z. Li, and C.P. Yuan, arXiv:1206.1344 [hep-ph].
[68] W. Skiba and D. Tucker-Smith, Phys. Rev. D **75**, 115010 (2007).
[69] B. Holdom, JHEP **08**, 069 (2007).
[70] L.G. Almeida et al., Phys. Rev. D **79**, 074017 (2009).
[71] A. Altheimer et al., J. Phys. G **39**, 063001 (2012).
[72] D.E. Acosta et al. (CDF Collaboration), Phys. Rev. D **71**, 112002 (2005).
[73] M.H. Seymour, Nucl.Phys. **B513**, 269 (1998).

[74] S.D. Ellis et al., JHEP **11**, 101 (2010).
[75] R. Kelley, M.D. Schwartz, and H.X. Zhu, arXiv: 1102.0561 [hep-ph].
[76] R. Kelley, M.D. Schwartz, R.M. Schabinger, and H.X. Zhu, Phys. Rev. D **86**, 054017 (2012).
[77] CMS Collaboration, Report CMS–PAS–QCD–10–014 (2010).
[78] M. Cacciari, G.P. Salam, and G. Soyez, JHEP **04**, 063 (2008).
[79] G. Buchalla, A. J. Buras, and M. E. Lautenbacher, Review of Modern Physics, **68**, 1125 (1996).
[80] G.P. Lepage and S.J. Brodsky, Phys. Lett. B **87**, 359 (1979); Phys. Rev. D **22**, 2157 (1980).
[81] A.V. Efremov and A.V. Radyushkin, Phys. Lett. B **94**, 245 (1980).
[82] V.L. Chernyak, A.R. Zhitnitsky, and V.G. Serbo, JETP Lett. **26**, 594 (1977).
[83] V.L. Chernyak and A.R. Zhitnitsky, Sov. J. Nucl. Phys. **31**, 544 (1980); Phys. Rep. **112**, 173 (1984).
[84] J.D. Bjorken, Nucl. Phys. Proc. Suppl. **11**, 325 (1989).
[85] M. Dugan and B. Grinstein, Phys. Lett. B **255**, 583 (1991).
[86] A.J. Buras and J.M. Gérard, Nucl. Phys. **B264**, 371 (1986); A.J. Buras, J.M. Gérard, and R. Rückl, Nucl. Phys. **B268**, 16 (1986).
[87] H.Y. Cheng, Phys. Lett. B **335**, 428 (1994); Z. Phys. C **69**, 647 (1996).
[88] J. Soares, Phys. Rev. D **51**, 3518 (1995); A.N. Kamal and A.B. Santra, Z. Phys. C **72**, 91 (1996); A.N. Kamal, A.B. Santra, and R.C. Verma, Phys. Rev. D **53**, 2506 (1996).
[89] M. Neubert, V. Rieckert, B. Stech, and Q.P. Xu, in *Heavy Flavours*, ed. A.J. Buras and M. Lindner (World Scientific, Singapore, 1992); M. Neubert and B. Stech, in *Heavy Flavours II*, ed. A.J. Buras and M. Lindner (World Scientific, Singapore, 1998), hep-ph/9705292.
[90] M. Bander, D. Silverman, and A. Soni, Phys. Rev. Lett. **43**, 242 (1979).
[91] A.P. Szczepaniak, E.M. Henley, and S.J. Brodsky, Phys. Lett. B **243**, 287 (1990).
[92] M. Beneke and M. Neubert Nucl. Phys. **B675**, 333 (2003).
[93] H.-n. Li, Phys. Rev. D **66**, 094010 (2002); H.-n. Li and K. Ukai, Phys. Lett. B **555**, 197 (2003).
[94] V.L. Chernyak and I.R. Zhitnitsky, Nucl. Phys. **B345**, 137 (1990).
[95] C.W. Bauer, D. Pirjol, and I.W. Stewart Phys. Rev. D **67**, 071502 (2003).
[96] C.W. Bauer, D. Pirjol, I.Z. Rothstein, and I.W. Stewart, Phys. Rev. D **70**, 054015 (2004).
[97] A. Khodjamirian, Th. Mannel, and B. Melic, Phys. Lett. B **571**, 75 (2003); Phys. Lett. B **572**, 171 (2003).
[98] Heavy Flavor Averaging Group, http://www.slac.stanford.edu/xorg/hfag.
[99] H.-n. Li and S. Mishima, Phys. Rev. D **73**, 114014 (2006).
[100] Y.Y. Charng and H.-n. Li, Phys. Rev. D **71**, 014036 (2005).
[101] T.N. Pham, hep-ph/0610063.
[102] M. Beneke and D. Yang, Nucl. Phys. **B736**, 34 (2006).
[103] M. Beneke and S. Jager, hep-ph/0512101.
[104] M. Beneke and S. Jager, Nucl. Phys. **B751**, 160 (2006).
[105] M. Beneke and M. Neubert, Nucl. Phys. **B675**, 333 (2003).
[106] C.H. Chen, Y.Y. Keum, and H.-n. Li, Phys. Rev. D **66**, 054013 (2002).
[107] H.-n. Li and S. Mishima, Phys. Rev. D **71**, 054025 (2005).
[108] Y. Grossman, Int. J. Mod. Phys. A **19**, 907 (2004).
[109] Y.D. Yang, R.M. Wang, and G.R. Lu, Phys. Rev. D **72**, 015009 (2005).
[110] P.K. Das and K.C. Yang, Phys. Rev. D **71**, 094002 (2005).
[111] C.H. Chen and C.Q. Geng, Phys. Rev. D **71**, 115004 (2005).
[112] C.S. Huang, P. Ko, X.H. Wu, and Y.D. Yang, Phys. Rev. D **73**, 034026 (2006).

[113] A.L. Kagan, Phys. Lett. B **601**, 151 (2004); hep-ph/0407076.

[114] M. Beneke, J. Rohrer, and D. Yang, Nucl. Phys. **B774**, 64 (2007).

[115] C.W. Bauer, D. Pirjol, I.Z. Rothstein, and I.W. Stewart, Phys. Rev. D **70**, 054015 (2004).

[116] P. Colangelo, F. De Fazio, and T.N. Pham, Phys. Lett. B **597**, 291 (2004).

[117] M. Ladisa, V. Laporta, G. Nardulli, and P. Santorelli, Phys. Rev. D **70**, 114025 (2004).

[118] H.Y. Cheng, C.K. Chua, and A. Soni, Phys. Rev. D **71**, 014030 (2005).

[119] W.S. Hou and M. Nagashima, hep-ph/0408007.

[120] M. Beneke, J. Rohrer, and D. Yang, Phys. Rev. Lett. **96**, 141801 (2006).

[121] H.-n. Li, Phys. Lett. B **622**, 63 (2005).

[122] A. Datta et al., Phys. Rev. D **76**, 034015 (2007).

[123] A. Ali et al., Phys. Rev. D **76**, 074018 (2007); Y. Li, C.D. Lu, Z.J. Xiao, and X.Q. Yu, Phys. Rev. D **70**, 034009 (2004).

[124] H.-n. Li, S. Mishima, and A.I. Sanda, Phys. Rev. D **72**, 114005 (2005).

[125] A.J. Buras, R. Fleischer, S. Recksiegel, and F. Schwab, Eur. Phys. J. C **45**, 701 (2006); R. Fleischer, S. Recksiegel, and F. Schwab, Eur. Phys. J. C **51**, 55 (2007).

[126] S. Baek and D. London, Phys. Lett. B **653**, 249 (2007).

[127] M. Gronau, arXiv:0706.2156 [hep-ph].

[128] A.V. Manohar and I.W. Stewart, Phys. Rev. D **76**, 074002 (2007).

Beyond the Standard Model

Mihoko M. Nojiri
Theory Center, IPNS, KEK, Tsukuba, Japan, and Kavli IPMU, The University of Tokyo, Kashiwa, Japan

Abstract
A Brief review on the physics beyond the Standard Model.

1 Quest of BSM

Although the standard model of elementary particles(SM) describes the high energy phenomena very well, particle physicists have been attracted by the physics beyond the Standard Model (BSM). There are very good reasons about this;

1. The SM Higgs sector is not natural.
2. There is no dark matter candidate in the SM.
3. Origin of three gauge interactions is not understood in the SM.
4. Cosmological observations suggest an inflation period in the early universe. The non-zero baryon number of our universe is not consistent with the inflation picture unless a new interaction is introduced.

The Higgs boson candidate was discovered recently. The study of the Higgs boson nature is extremely important for the BSM study.

The Higgs boson is a spin 0 particle, and the structure of the radiative correction is quite different from those of fermions and gauge bosons. The correction of the Higgs boson mass is proportional to the cut-off scale, called “quadratic divergence". If the cut-off scale is high, the correction becomes unacceptably large compared with the on-shell mass of the Higgs boson. This is often called a “fine turning problem". Note that such quadratic divergence does not appear in the radiative correction to the fermion and gauge boson masses. They are protected by the chiral and gauge symmetries, respectively.

The problem can be solved if there are an intermediate scale where new particles appears, and the radiative correction from the new particles compensates the SM radiative correction. The scale is probably much less than $\mathcal{O}(100)$ TeV, where the ratio between the SM radiative correction and the Higgs vev is more than 1000. The turning of the factor 1000 may sound unnatural, but it is much better than the scale among other parameters, such as Planck scale to the order of electroweak symmetry breaking, or the large difference among Yukawa couplings.

An idea to introduce a new particle that couples to the Higgs boson to cancel one loop level quadratic correction, is not successful, because such accidental cancellation does not hold all order in the perturbation theory. One needs new symmetry to cancel the quadratic divergence in the SM by a new physics contribution. The known ideas to achieve the reduction of quadratic divergence are the following;

1. **Supersymmetry**: Extend the SM so that the theory has “supersymmetry". Supersymmetry is the symmetry between bosons and fermions, which allows the divergence of Higgs boson mass controlled by "chiral symmetry" of fermions. Due to the cancellation among various diagrams involving SM particles and their superpartners (SUSY particles), there are no quadratic divergence to the Higgs bosons mass in this theory.
2. **Dynamical symmetry breaking**: In this theory, a new strong interaction causes the spontaneous gauge symmetry breaking of the SM. The Higgs doublet is a Nambu-Goldstone boson of the symmetry breaking and bound states of fermions charged under the strong interaction, corresponding

to the pions in the QCD. The Higgs boson does not exist above the symmetry breaking scale, so there are no problem of quadratic divergence.

3. **Extra dimension** Although we recognize that we live in the four dimensional space-time, we might live in more than the five dimension space time where the extra dimensions are compactified. The true Planck scale may be much closer to electroweak scale in such a theory, or the fundamental parameters in the Higgs sector is of the order of Planck scale in the higher dimensional theory but looks small in the effective four dimensional theory. In some class of the model the Higgs boson may be a part of gauge boson in the 5th dimension so that the divergence of the Higgs mass parameters is controlled by the gauge symmetry.

Those models are constrained strongly by precision measurements. Currently there are no measurements with significant deviation from the SM predictions. In the SM theory, one can predict various observables from a few fundamental parameters: the gauge couplings $g_i (i = 1, 2, 3)$, and the Higgs vacuum expectation value (vev) v. By measuring the deviations from the SM predictions, we can set constrains on the new physics. Especially, the S and T parameters which parametrize the new physics contributions to the gauge two point functions are sensitive to all particles that couple to the gauge bosons. Measurements of flavor changing neutral current (FCNC) constrain the existence of flavor off-diagonal interactions. Very precisely measured parameters sometimes exhibit significant deviations from the SM predictions. Currently muon anomalous magnetic moment deviates from the SM prediction by more than 3 σ. It is sensitive to the new physics that couples to muon.

The quadratic divergence of the Higgs sector exists if the divergence is estimated by the momentum cut off Λ, the upper bound of the various loop integral appearing in the radiative correction in the mass. We have to keep it in mind that the quadratic divergence does not depend on the external momentum, therefore it is a regularization dependent object. Especially in dimensional regularization, quadratic divergence is trivially zero. Then, is there any reason that we should take the fine turning problem seriously?

The fine turning argument based on momentum cut-off is justified in the case that the theory has large symmetry at some higher energy scale. For example, in the supersymmetric model, the regularization must respect to supersymmetry and one cannot subtract all quadratic divergence. To this end, the Higgs sector receives radiative corrections proportional to the SUSY scale (superpartner mass scale) under correct regularization. In the limit that superpartners are much heavier than SM particles, the low energy theory looks like the SM with the momentum cutoff at the SUSY scale. Fine turning arguments hold for the theories with an intermediated scale above which a new symmetry emerges.

There is another indication of the existence of new physics between the weak scale and the Planck scale. We may consider the Higgs potential at large field value in the SM and study the stability. The potential is a function of the top and Higgs masses, and current top and Higgs mass measurements favor metastable Higgs potential. There is not any reason that the Higgs vev should fall in such a metastable point, and this also suggests that additional particles that couple to the Higgs sector change the shape of the potential.

Another strong indication of new physics is the existence of dark matter in our Universe. Global fit of the cosmological observation favors the existence of stable, neutral particle, dark matter, which accounts for 27% of the total energy of our Universe. The existence of the dark matter is also confirmed by various observations of the stellar objects. Rotation curve of the stars of the galaxy indicates that galaxies are dominated by the non-luminous component. The is also a technique to measure the matters extended beyond the galaxy scale using gravitational lensing.

Our universe is 1.38×10^{10} years old, roughly 10^{17} s $\leftrightarrow 10^{-43}$ GeV^{-1}. The dark matter life time must be at least of the oder of the age of the Universe to remain in the current Universe. [1] On the other

[1] In order to avoid the constraints coming from cosmic ray observations, the lifetime of the dark matter in our Universe must be significantly longer than the age of the Universe.

hand, a particle with mass m (GeV) with interaction suppressed by $1/M_{pl}$ has a decay width of order of $g^2(m/1\ \mathrm{GeV})^3 10^{-38}$ GeV. Namely the lifetime, $\tau \sim g^{-2}10^{14}\mathrm{s}/(\mathrm{m}/1\ \mathrm{GeV})^3$, would be much shorter than the life of our Universe ($\sim 4.3\times 10^{17}$ s), where g is the coupling of the decay vertex. To account for the lifetime of the dark matter in our universe, its decay must be very strongly suppressed, or forbidden.

For the case of the SM particles, existence of stable particles is ensured by the symmetry. Electron is the lightest charged particle and electronic charge is conserved by the gauge symmetry. Proton is the lightest bound state of quarks. There are no interaction to break proton in the SM, because number of quark is conserved for interaction with the gauge bosons or the Higgs boson, and direct interaction with electron is forbidden by the gauge symmetry. It is possible to conserve the Baryon number 1/3 to the quarks in the SM, and this reflects the fact that proton is stable. To consider the particle model involving the stable (or long-lived) dark matter, we must introduce new symmetry to protect the dark matter from decaying.

Another puzzle of the SM is the hyper-charge assignments of the fermions. In the first glance, it is not easy to find the rules to assign the charge to the SM matters. But, it fits very nicely to the representation of a $SU(5)$ group, where $SU(3)\times SU(2)\times U(1)$ generators are embedded as

$$T^a_{SU(3)} = \begin{pmatrix} \lambda^a & 0 \\ 0 & 0 \end{pmatrix} \quad T^i_{SU(2)} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma^i \end{pmatrix} \quad T_{U(1)} = \begin{pmatrix} -\frac{1}{3}\mathbf{1}_3 & 0 \\ 0 & \frac{1}{2}\mathbf{1}_2 \end{pmatrix}. \tag{1}$$

Here, λ^a and σ^i are the $SU(3)$ and $SU(2)$ generators, $\mathbf{1}_3$ and $\mathbf{1}_2$ are 3×3 or 2×2 unit matrix, and $T_{SU(3)}, T_{SU(2)}, T_{U(1)}$ satisfy the commutation relations of $SU(3)$, $SU(2)$, and $U(1)$ generators. Under this generator assignment, $\mathbf{5}^*$ and $\mathbf{10}$ representations of $SU(5)$ have a charge assignment as

$$\mathbf{5}^* = \begin{pmatrix} (3^*,1)_{1/3} \\ (1,2)_{-1/2} \end{pmatrix}, \tag{2}$$

while $\mathbf{10}$ representation is decomposed into $(3,2)_{1/6}\oplus(3^*,1)_{-2/3}\oplus(1,1)_1$ which reside in the 5×5 antisymmetric matrix as

$$\mathbf{10} = \begin{pmatrix} (3^*,1)_{-2/3} & (3,2)_{1/6} \\ * & (1,1)_1 \end{pmatrix}. \tag{3}$$

This suggests that $SU(3)\times SU(2)\times U(1)$ symmetry of the SM can be unified into the $SU(5)$ gauge symmetry. To realize this, the SM three gauge couplings must unify at the short distance, so that the $SU(5)$ symmetry is recovered above that scale. The gauge couplings at the short distance is calculated by utilizing the SM renormalization group equations from the low energy inputs. They do not unify for the particle content of the SM, therefore to realize the idea of GUT, new set of particles are needed. We will see a successful gauge coupling unification is realized in the Supersymmetric model in the next section.

2 Supersymmetry

Supersymmety is the symmetry exchanging bosons into fermion, and fermions into bosons. The generators of the supersymmetric transformation satisfy the following anti-commutation relations

$$\left\{Q^\alpha, \bar{Q}_{\dot\beta}\right\} = 2\sigma^\mu_{\alpha,\dot\beta}P_\mu \tag{4}$$

Here Q is a spin $1/2$ and mass dimension $1/2$ operator and α and $\dot\beta$ $(=1,2)$ are the spin indices of chiral and anti-chiral fermions, and $\sigma^\mu = (1,\sigma^i)$ is the Pauli matrices.

This anti-commutation relation can be reduced for any massive eigenstate $|a\rangle$ by taking the rest frame $P^\mu|a\rangle = m_a\delta_{0\mu}|a\rangle$ as follows:

$$\left\{Q^\alpha, \bar{Q}_{\dot\beta}\right\} = 2\delta_{\alpha,\dot\beta}m_a. \tag{5}$$

Table 1: Particle content of the Minimal Supersymmetric Standard Model.

represenations	quark	squark
$(3,2)_{1/6}$	$q_L = (u,d)_L$	$\tilde{q}_L = (\tilde{u}_L, \tilde{d}_L)$
$(3^*,1)_{-2/3}$	u_R^c	$(\tilde{u}_R)^c$
$(3^*,1)_{1/3}$	$(d_R)^c$	$(\tilde{d}_R)^c$
	lepton	slepton
$(1,2)_{1/2}$	$l_L = (\nu, e)_L$	$\tilde{q}_L = (\tilde{\nu}_L, \tilde{e}_L)$
$(1,1)_1$	$(e_R)^c$	$(\tilde{e}_R)^c$
	Higgsino	Higgs
$(1,2)_{-1/2}$	$(\tilde{H}_1^0, \tilde{H}_1^-)$	(H_1^0, H_1^-)
$(1,2)_{1/2}$	$(\tilde{H}_2^+, \tilde{H}_2^0)$	(H_2^+, H_2^0)
	spin 1/2	spin 1
$(8,1)_0$	$\tilde{G}$ (gluino)	G^μ
$(1,3)_0$	$\tilde{W}$ (wino)	W^μ
$(1,1)_0$	$\tilde{B}$ (bino)	B^μ

The relation is same as that of a two-fermion system in quantum mechanics. One can construct an irreducible representation of this algebra starting from a state which annihilates any $\bar{Q}_i$. Suppose the state is spin 0, $|0\rangle$, all possible states are generated as follows;

$$|0\rangle \to Q_1|0\rangle, Q_2|0\rangle \to Q_1Q_2|0\rangle. \tag{6}$$

Because $Q_1Q_1 = Q_2Q_2 = 0$, no more state can be obtained by multiplying the generator Q_i. Two spin 0 states and two spin $1/2$ states are obtained. These states form a SUSY multiplet, and the spin 0 states are the superpartners of the spin $1/2$ states and vise versa. Because this multiplet contains spin 1/2 states, we can regard this as a matter multiplet.

Starting from a spin 1/2 state annihilating $\bar{Q}$ one gets two spin $1/2$ fermion states, a spin 1 massive bosonic states and a spin 0 bosonic state, namely 4 fermion degrees of freedom and 4 bosonic degrees of freedom. This may be regarded as two chiral fermions, one massive gauge boson and one massive Higgs boson. Repeating similar analysis to the massless particles, one obtains states with helicity $h = \lambda$ and $\lambda + 1/2$. If $\lambda = 1/2$, a massless gauge boson and its superpartner fermion make a supersymmetric multiplet. The number of bosonic degrees of freedom is the same as that of fermionic degrees of freedom in this theory.

All states in the above multiplet have the same mass, which looks irrelevant for describing real particles, but it is known that such mass degeneracy is removed by spontaneous supersymmetry breaking. Supersymmetry breaking is discussed in the next section.

The minimal supersymmetric standard model (MSSM) is an extension of the SM that has a supersymmetry in the limit where all particle masses are ignored. The model is thought to be an effective theory of a fully supersymmetic theory. Due to the spontaneous supersymmetry breaking of the full theory, the superparters of the SM particles receive a mass much higher than the SM particles. A superpartner of a fermion is called sfermion and it is a spin 0 particle. A superpartner of a gauge boson is called gaugino and has spin $1/2$. A Higgs boson superpartner is called a higgsino and has spin $1/2$. The particle content of the MSSM is given in Table 2. The SM particles and their superpartners have same charge, because the generator of supersymmetric transformation Q commutes with the SM $SU(3) \times SU(2) \times U(1)$ transformation. The number of Higgs doublets is two in the MSSM because one should add two Higgsinos, chiral fermions with charge $(1,2)_{\pm 1/2}$ in the SM because of a condition of anomaly cancellation.

As one can see from Table 2, the number of particles are doubled in the MSSM. The supersymmetry specifies all dimensionless couplings of interactions of new particles, such as four point interaction of

scalers and Yukawa couplings, while mass parameters of superpartners are undetermined. To understand the coupling relations, one needs to understand the supersymmetric field theory. In this lecture, I do not have enough time to talk about it in detail, so I just sketch the important elements.

Fields in the same supersymmetric matter multiplet can be arranged in a "chiral superfield" which is a function of coordinate x, θ and $\bar{\theta}$ a grassmanian Lorentz spinors with mass dimension $-1/2$,

$$\Phi(x,\theta,\bar{\theta}) = \phi(y) + \sqrt{2}\psi(y)\theta + F(y)\theta\theta, \tag{7}$$

where $y^\mu = x^\mu - i\theta\sigma^\mu\theta$. Note that by redefining the coordinate from x to y, Φ becomes a function of y and θ, and $\bar{\theta}$ does not appear. There are only three fields ϕ, F and ψ appearing as the component fields of Φ. When θ is zero, $\Phi(x) = \phi(x)$, therefore Φ is an extension of the scalar field of non-supersymmetric theory. On the other hand, $\Phi(y,\theta)$ represents both fermonic and bosonic fields simultaneously.

Φ is dimension 1, so that $\dim(\phi) = 1$ and $\dim(\psi) = 3/2$. F is then spin 0 and dim 2 field. The only $dim < 4$ kinetic term of F is FF^*, therefore F is not dynamical. The product of a chiral superfield is also a chiral superfield depending only y and θ. On the other hand, $\Phi\bar{\Phi}'$ is not a chiral superfield as it has the terms proportional to $\bar{\theta}$.

Just as operator P^μ,translation in coordinate space x is expressed as $\partial/\partial x$, supersymmetric transformation Q is a translation in the θ and $\bar{\theta}$ space. Namely, in the coordinate representation it is expressed as

$$S_\alpha = \frac{\partial}{\partial\theta^\alpha} + i(\sigma^\mu\partial_\mu\bar{\theta}). \tag{8}$$

The second term is needed to satisfy the SUSY algebra give in Eq. 4. With this transformation, each field transform as

$$\begin{aligned} \delta_{SUSY}\phi &= \sqrt{2}\alpha\psi, \\ \delta_{SUSY}\psi &= -i\sqrt{\partial}_\mu\sigma^\mu\phi\bar{\alpha} + \sqrt{2}F\alpha, \\ \delta_{SUSY}F &= -i\sqrt{2}\bar{\alpha}\partial_\mu\bar{\sigma}^\mu\psi, \end{aligned} \tag{9}$$

where α and $\bar{\alpha}$ are transformation parameters. Under this transformation kinetic term

$$\mathcal{L}_{kin} = \partial_\mu\phi\partial^\mu\phi^* + i\bar{\psi}\sigma^\mu\partial_\mu\psi + F^*F \tag{10}$$

is invariant.

There are a few things worth paying attention. First The $\delta_{SUSY}F$ is total derivative. Because the product of chiral superfields is also a superfield, the $\theta\theta$ component $\mathcal{F}$ transforms as $\mathcal{F} = \partial_\mu J^\mu$, namely $\mathcal{F}$ can be interaction terms which are invariant under supersymmetric transformation. For example, $\Phi_1\Phi_2\Phi_3$ gives F term

$$L_{Yukawa} = F_1\phi_2\phi_3 + F_2\phi_1\phi_3 + F_3\phi_1\phi_2 - \psi_1\psi_2\phi_3 - \psi_2\psi_3\phi_1 - \psi_3\psi_1\phi_2. \tag{11}$$

The interaction contains Yukawa interaction term $y_{ijk}\psi_i\psi_j\phi_k$ which is symmetric under the exchange of i,j,k, and also the scalar potential terms proportional to $y_{ijk}F_i\phi_j\phi_k$ Combined with kinetic term FF^*, interactions of four point scalar fields proportional to y^2 is generated. The similar relations also holds for supersymmetric gauge interactions. The interaction between gaugino-fermion- sfermion is proportional the gauge coupling g, and there are scalar four point interactions proportional to g^2. While many scalar and fermion partners are introduced, there are no new dimensionless coupling introduced.

In addition to the F term, $\theta\theta\bar{\theta}\bar{\theta}$ term of general field product, $\mathcal{D}$ is supersymmetric. For example, supersymmetric kinetic term is $\theta\theta\bar{\theta}\bar{\theta}$ term of $\Phi\bar{\Phi}$.

We now address some important features of supersymmetric models.

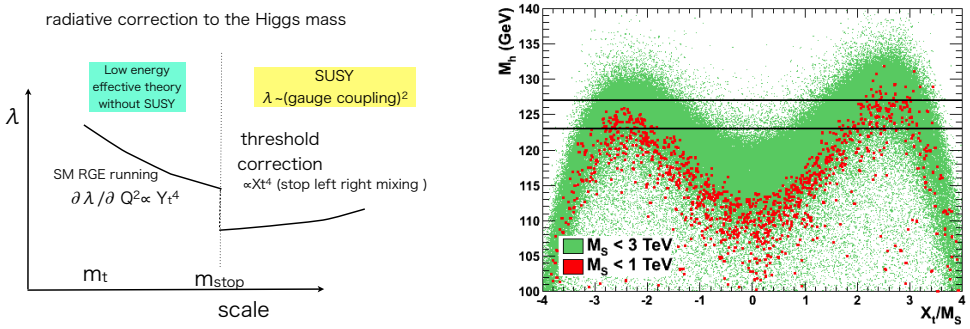


Fig. 1: Left: the running of Higgs four point coupling changes at the scale of m_{stop}. Right: Maximal value of the Higgs boson mass as a function of X_t/M_{SUSY} when all the other parameter are scanned. From arXive 1311.0720.

– There are no quadratic divergence in the theory. The quadratic divergence coming from the top loop is canceled by the stop loop generated by the Higgs-Higgs-stop-stop four point interaction. Both of them are proportional to y^2. The Higgs four point coupling is proportional to the square of the gauge coupling, and quadratic divergence arising from the diagram is canceled by the gauge and gaugino-higgsino loops. This is because scalar particles are now in a same multiplet with the fermion, and the mass of the fermion is only logarithmically divergent. The fine-turning in the Higgs sector is now significantly reduced.

– Because the Higgs four point coupling is a gauge coupling, the Planck scale Higgs four point coupling is always positive, therefore significantly less in danger of running into metastable vacuum At low energy the Higgs mass is upper bounded by the Z boson mass in tree level, and radiative corrections proportional to the $(mt^4/m_W^2)\log(m_{\tilde{t}}/m_t)$ appear in the Higgs boson mass formulae. This correction is interpreted as the running of the Higgs boson four point coupling from the stop mass scale to the top mass scale under the SM renormalization group equation, because below the stop mass scale, the theory is effectively the SM. In addition there are contribution proportional to the fourth power of stop left-right mixing X_t. See Fig. 1 (left) for the RGE interpretation of the radiative corrections to the Higgs mass. In this theory, the Higgs boson mass is calculated from the scalar top mass and its mixing, therefore the SUSY scale is predicted from the Higgs boson mass. In other words, the measured Higgs boson mass gives a strong constraint to the SUSY mass scale and mixing. See Fig. 1 (right).

– In the SM, one cannot write an interaction violating baryon and lepton numbers due to the gauge invariance. This is no longer true because Higgsino and lepton doublets have same quantum numbers. The product of superfields W whose $\theta\theta$ terms is the SM Yukawa interactions

$$W = -y_e H_1 \cdot E^c L - y_d H_1 \cdot D^c Q - y_u H_2 \cdot U^c Q - \mu H_1 \cdot H_2, \tag{12}$$

where $Q = \tilde{q}_L + \theta q_L ...$, $U^c = \tilde{u}^c_R + \theta u^c_R .....$ are the superfields whose bosonic component is a sfermion and a fermionic component is quarks or leptons. However, $\theta\theta$ term of W'

$$W' = \epsilon_L LLE^c + \epsilon_{BL} LQD^c + \epsilon_B U^c D^c D^c + \epsilon_{LH} LH_2 \tag{13}$$

is not forbidden by the gauge symmetry, because H_1 and L have a same quantum numbers, and $UDD = \epsilon_{abc} U^a D^b D^c$ is a gauge singlet. The interactions violate lepton and/or baryon numbers and should be forbidden.
The symmetry that forbids L and B violating terms is called the conserved R-parity. In the MSSM R-parity may be assigned to the superfield and coordinate θ as follows,

$$R(L) = R(E) = R(Q) = R(U) = R(D) = -1, R(H) = 1, R(\theta) = -1. \tag{14}$$

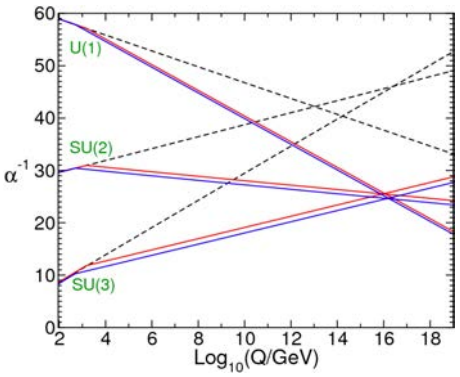


Fig. 2: Two loop renormalization group evolution of the gauge couplings in the SM (dashed lines) and MSSM (solid lines) from "A Supersymmetry primer" hep-ph/9709356.

In this assignment, all the SM particles have $R = 1$ and all superpartners have $R = -1$, and $R(W|_{\theta\theta}) = 1$, and $R(W'|_{\theta\theta}) = -1$. The interaction term from W multiplicatively conserves R parity, namely, product of R parity of all particles involved in a vertex is one. Namely, $R = -1$ particle decays into the final states which contains odd number of $R = -1$ particles. If two $R = 1$ particle collides, the final state contains even number of $R = -1$ particles. By requiring multiplicatively conserved R parity, the lightest supersymmetric particle (LSP) becomes stable. The LSP can be a dark matter candidate.

– Gauge coupling: In the supersymmetric model, the number of particles is doubled and running of the gauge couplings would be modified above the SUSY particle mass scale. The gauge couplings unify at the GUT scale much better than that of the SM as can be seen in Fig. 2. This means "supersymmetric GUT" is consistent with experimental data, though there are still some fine turning issues when we consider the Higgs sector violating GUT symmetry.

3 Origin of SUSY breaking

As we have mentioned already, the MSSM is not a complete theory, because it requires a mechanism to break the supersymmetry somewhere outside the MSSM. A general set up of the SUSY breaking models are the following; there are hidden sector H, and fields Z_i in the sector H break the supersymmetry spontaneously. This hidden sector couples to our sector indirectly though a messenger sector. The particles in the messenger sector have a mass scale M.

The spontaneous symmetry breaking is realized for the vacuums which do not annihilate with the supersymmetric generator Q and $\bar{Q}$. If such a vacuum exists, there are some fermions ψ whose supersymmetric transformation $\delta_{SUSY}\psi = \{Q, \psi\}$ has non-zero vev, namely $\langle 0|\delta_{SUSY}\phi|0\rangle = -\sqrt{2}\langle 0|F|0\rangle \neq 0$. Some of the superfields in the Hidden section must have non-zero F terms in our setup.

If F term of Z has non zero vev, $\langle Z\rangle = \langle F_Z\rangle\theta\theta$, various mass terms are induced in the low energy effectively. A simple example is $\theta\theta\bar{\theta}\bar{\theta}$ term of $Z\bar{Z}\Phi\bar{\Phi}/M^2$, which may be induced through the messenger interactions. After the symmetry breaking the term $(\langle F\rangle^2/M^2)\phi\phi^*$ is the effective SUSY breaking mass term of the scaler boson ϕ.

There are already severe constraints to the interaction of the messenger sector to the MSSM sector. These constrains come from the flavor changing neutral currents such as K^0-$\bar{K}^0$ mixing. The constraints typically require

$$\left[\frac{10\text{TeV}}{m_{\tilde{q},\tilde{g}}}\right]^2 \left[\frac{\Delta m^2_{\tilde{q}_{12}}/m^2}{0.1}\right]^2 < 1, \tag{15}$$

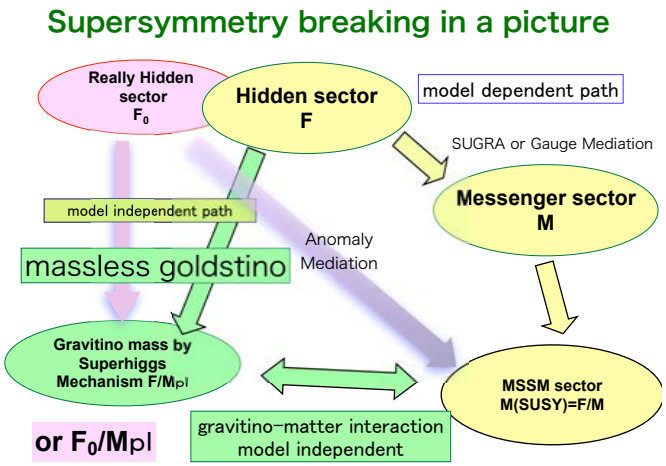


Fig. 3: Relation between the MSSM sector and SUSY breaking sector.

where $m^2_{\tilde{q}_{12}}$ is a mixing parameter of the first and second generation squark, and m^2 is diagonal squark masses. The SUSY breaking sector H therefore must couple to the MSSM matter sector universally.

Several mechanisms have been proposed to assure the universality of the soft scalar masses. The supergravity model uses the gravity interaction as the messenger mechanism, on the other hand, gauge mediation models uses some vector-like matters charged under the SM gauge groups as the messenger fields. Even if there are no direct couplings between the MSSM and SUSY breaking sectors, there are mediation mechanism through the superconformal anomaly, and the model utilizing this is called anomaly mediation model.

It is difficult to access the HIdden sector directly. The SUSY breaking of the total theory F_0 and mass of the gravitino(super partner of graviton) $m_{3/2}$ is related as $m_{3/2} = F_0/M_{pl}$. The gravitino could be the LSP, in that case the next lightest SUSY particle(NLSP) is long-lived. The NLSP can be detected directly at the collider, the decay lifetime provide the information of hidden sector SUSY breaking. If gravitino is not the LSP, the gravitino can be long-lived and may have impact on big-bang neucleosynthesis. See Fig. 3.

The mediation mechanism sets the sparticle mass parameters at the mediation scale, and on-shell masses of the SUSY particles are obtained by running the RGE equation of the masses down to the low energy scale. If the boundary condition is universal at M_{GUT}, squark and gluino masses are much heavier than those of electroweakly interacting superpartners such as sleptons, wino, bino and Higgsinos. The square of Higgs mass parameter is driven to be negative at the weak scale, and Higgsino mass parameter μ compensates it so that the Higgs vev is the correct value. The cancellation between μ and SUSY breaking parameters at the weak scale is a measure of the fine turning in the Higgs sector. See Fig. 4.

4 Collider search of supersymmetric particles

So far, a proton-proton collider at CERN, the Large Hadron Collider (LHC), has collected ~ 30 fb^{-1} of integrated luminosity for each experiment at 7 to 8 TeV. It will start operation again from 2015 aiming for 300 fb^{-1} at 13 TeV.

A proton is a composite particle and quarks and gluons in the proton are the elementary particles that are involved in the high energy scatting process. The momentum of the quarks and gluons are parallel to the beam direction but the absolute values are not fixed. Therefore the collision system is boosted to one of the beam directions. The production cross section is generally the highest near the threshold. It

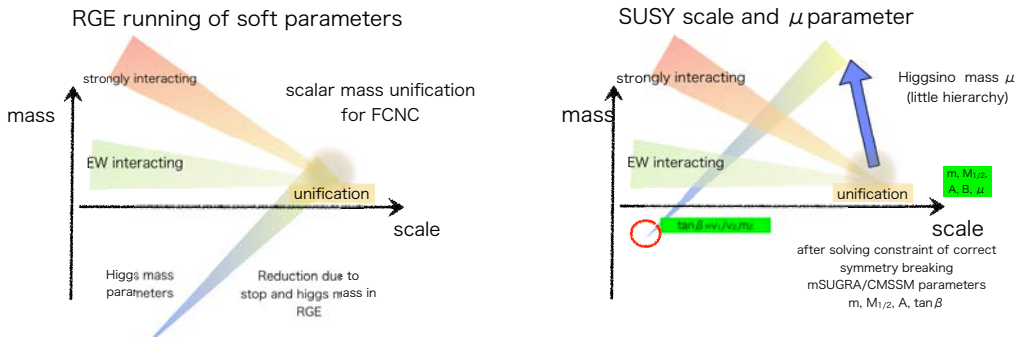


Fig. 4: Relation between the MSSM sector and SUSY breaking sector.

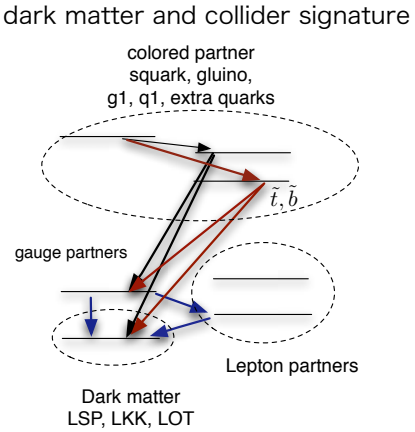


Fig. 5: The decay pattern of squark and gluino produced at the LHC, and particles emitted from the cascade decay chain. The particles in the Little Higgs model with T parity or universal extra dimension model may also give a similar signature.

reduces gradually with the increase of the parton collision energy $\sqrt{s}$. The quarks and gluons in the final state are fragmented and hadronized into hadrons, forming the jets. Electroweakly interacting particles W, Z, γ, leptons and neutrinos are also produced from various production processes.

Colored superpaticles are copiously produced at the hadron collider. Due to the conserved R-parity of the MSSM, superpartners are produced in pairs, each superpartner decays to the final state involving another superpartner, and at the end of the cascade decay, the LSP appears. The LSP is stable. Due to the cosmological constraints, it is neutral and color-singlet, and escapes detection. If the mass difference between the superpartners are large, the decay product tends to have high p_T. In such a case, the LSP, which cannot be detected directly, is also relativistic (See Fig. 5). The sum of LSP momentum transverse to the beam direction is balanced against other visible particles. Namely, significant missing transverse momentum $\mathbf{P}_{Tmiss}$ defined as

$$\mathbf{P}_{Tmiss} = -\sum_i \mathbf{p_T}^i_{\rm jet} + \sum_j \mathbf{p_T}^j_{\rm l}, \tag{16}$$

is a signature of SUSY particle production. Another important quantity is the sum of absolute values of

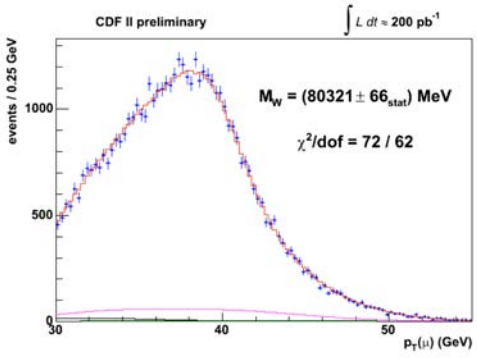


Fig. 6: Distribution of p_T^μ from the W boson decay measured at CDF experiment at Tevatron.

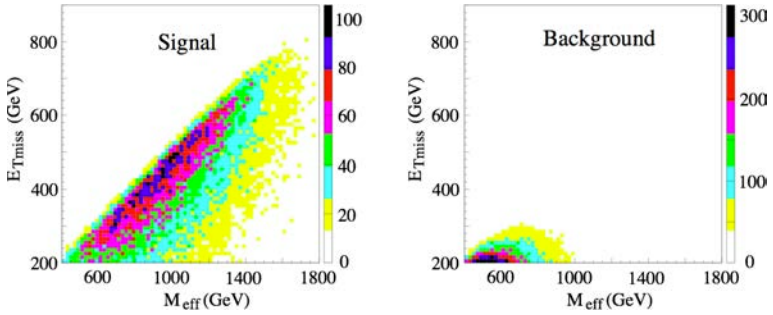


Fig. 7: Distribution of m_{eff} and E_{Tmiss} of the top partner pair production at the LHC followed by the decay into top and stable neutral gauge boson (left) compared with the $t\bar{t}$ distribution.

the transverse momentum

$$H_T = \sum_i p^i_{\rm Tjet} + \sum_j p^j_{\rm Tl}, \tag{17}$$

or the effective mass

$$m_{eff} = \sum_i p^i_{Tjet} + \sum_j p^j_{Tl} + E_{Tmiss}, \tag{18}$$

where E_{Tmiss} is the absolute value of missing transverse momentum.

The $m_{\rm eff}$ distribution peaks at the sum of the produced particles at the hard process. To observe this fact, let us first consider the p_T distribution of leptons from W boson decay produced at CDF experiment at Tevatron, a $p\bar{p}$ collider at 1.8 TeV. The distribution peaks at 40 GeV, which is a half of the W boson mass. See Fig. 6. The feature is easily understood when we calculate the p_T distribution of spherically decaying W boson boosted to the beam direction,

$$f(x)dx = \frac{2}{\sqrt{1-x^2}}dx, \tag{19}$$

where $p_T = (m_W/2)\sin\theta = xm_W/2$; The distribution strongly peaks at $p_T = m_W/2$ ($\sin\theta = 1$) and the structure remains even though W bosons are boosted transversely in the realistic situation, because the production cross section is largest near the threshold. The fact applies to all production processes at the hadron collider; the sum of the p_T of the decay products peaks near the parent's mass. When heavy particles are produced in pairs, the sum of the p_T of the decay products peaks at the sum of the produced particle masses.

Fig. 7 compares the distributions of T_-T_- and $t\bar{t}$ pair productions. Here a hypothetical particles T_- is assumed to decay into t and B_H, and B_H is a neutral stable massive $U(1)$ gauge boson. The signal

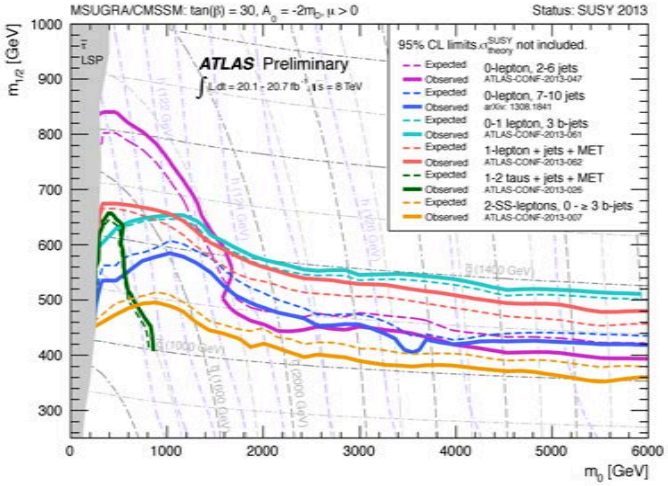


Fig. 8: Latest mass limit of the MSSM squarks and gluinos shown as a function of GUT scale gaugino mass and scaler mass. Presented in SUSY2013.

contains $t\bar{t}$ and existence of two B_H's is observed by the missing transverse momentum of the events, namely, the signal is similar to that of superpartner pair production. The signal production cross section is $\mathcal{O}(1)$ pb, while the $t\bar{t}$ production cross section is huge at the LHC, around 800 pb. If the distribution overlaps significantly, the signal is very difficult to be observed. However, the signal m_{eff} distribution peaks around 1 TeV and missing momentum as close as half of the M_{eff}, while the background peaks around $m_{eff} \sim 400$ GeV and $E_{Tmiss} \ll M_{eff}/2$. Because of this distribution differences, the T_{-} signature with the production cross section much less than 1 pb may be observed at the LHC.

So far we have been talking about "inclusive" quantity. They are defined using all objects in an event. We may also select jets or leptons with special features and use kinematical information to separate signals and backgrounds. Let us consider events with one lepton and some missing momentum. The event with one lepton + multiple jets + missing momentum is an important signature of superpartner production. However, events involving W boson also produce such signatures. However, the events with W boson can be reduced significantly if we require that m_T of a lepton and missing p_T is above 100 GeV where m_T is defined as

$$m_T = \sqrt{2p_T^l E^{Tmiss}}(1 - \cos(\Delta\phi(l, p_T)) \tag{20}$$

The cut significantly reduces the background from the W boson production to the SUSY process.

The current bound of the SUSY process is obtained after successful reduction of background using the above kinematical variable. The understanding of background distribution is quite important, especially the cross section of W,Z, $t\bar{t}$ with multiple jets must be correctly calculated. The techniques to obtain multiple jets amplitudes with parton shower has been established only this century, and current SUSY searches at the LHC is benefitted by those techniques greatly. The current limit typically excludes squark with mass 1.8 TeV and gluino with mass less than 1.4 TeV, if the mass splitting between the LSP and colored SUSY particles are large enough. See Fig. refsusylimit for the latest limits.

5 Dynamical symmetry breaking and BSM

Supersymmetry is not a unique solution of the hierarchy problem. Another important class of solutions is dynamical symmetry breaking models. When a global symmetry is broken spontaneously, a massless scalar modes (Nambu-Goldstone boson) appears, even if the theory does not have an elementary Higgs boson. An important example is chiral symmetry breaking in QCD. The QCD Lagrangian has $SU(2)_L \times SU(2)_R$ symmetry when quark masses are ignored. The symmetry is spontaneously broken to $SU(2)_V$

dynamically, and the Goldstone boson of the symmetry breaking are pions $\pi \sim \bar{q}_i\gamma_5 q'$, and $\langle\bar{q}q\rangle$ has non-zero vev.

The pion has the same charge as the Goldstone boson in the Higgs sector. Therefore, it is natural to consider scale up of the mechanism. The model involves a set of new quarks Q with EW charges, but couple to different asymptotic free gauge interactions whose couplings blow up at the scale of EW symmetry breaking. If $\bar{Q}Q$ condense, the light $\bar{Q}\gamma_5 Q$ states work as the Goldstone bosons of the EW symmetry breaking. This class of the model called Technicolor model. The model has no quadratic divergence because the massless bound states only appear in the low energy effective theory.

This is an interesting and beautiful idea, but is not consistent with precision EW observations. At LEP, gauge boson two point functions were precisely measured. Especially the parameter called S, receives non decoupling contribution from $SU(2)$ doublets Q which is colored in the new strong interactions, and also necessary charged under $SU(2) \times U(1)$ symmetry in the SM to break the gauge symmetry. Their contribution appears constructively to the gauge two point functions, and therefore the model is tightly constrained. In addition, these models tend to predict a heavy Higgs boson inconsistent with the data.

Another class of models called "composite Higgs models" allows a Higgs boson which is light but non-elementary. In these models, the Higgs doublet itself is a pseudo Goldstone boson of some dynamical symmetry breaking. Though the mechanism of dynamical symmetry breaking is not specified, the smallness of the mass of the Higgs boson is thought to be ensured by the global symmetry of the theory. The model requires extension of the top sector because the top Yukawa coupling violates the desired global symmetry strongly. The extended top sector is a target of extensive ATLAS and CMS searches.

6 Extra dimension models

In the Extra dimension models the space has more than three dimensions but the additional space dimension is compactified with a small size R so that we could not recognized it easily. When the extra dimension is flat, the fields in the extra dimension may satisfy the periodic boundary condition such as

$$\phi(x, y) = \phi(x, y + R), \tag{21}$$

where x represents four dimensional space time, while y is the fifth dimension. Under this boundary condition, the wave function is expressed as

$$\psi(x, y) = \psi'(x) \exp(ip_5 y), \tag{22}$$

where $p_5 R = 2\pi n$ (n is an integer). This leads to an equation of motion of a free particle propagating in the the fifth dimension,

$$E_n^2 = p^2 + p_5^2 = p^2 + (2\pi)^2 \left(\frac{n}{R}\right)^2. \tag{23}$$

Namely, the model predicts an infinite tower of particles of the four dimensional effective theory, which corresponds to different values of the discrete momenta in the fifth direction.

The coupling of the fifth dimension related with the couplings in the four dimensional effective theory in non-trivial manner. A simple example is the gauge coupling of the fifth dimensional theory and the four dimensional effective theory,

$$\int d^4x dx_5 \frac{1}{g_5^2} F_{\mu\nu} F^{\mu\nu} \to \int d^4x \frac{1}{g_4^2} F_{\mu\nu} F^{\mu\nu}, \tag{24}$$

where $g_4 = g_5/\sqrt{R}$. Larger the size of the fifth dimension is, g_4 becomes small. This is also true for gravitational interactions. The four dimensional gravitational interaction may be small because the size

of extra dimension is large. The Large extra dimension model tried to solve the fine tuning problem by making true Phanck scale in the higher dimensional theory much smaller than the M_{pl}.

The extra dimension may not be flat. In the RS model, the fifth dimension has non-trivial metric as follows:

$$ds^2 = e^{-2\sigma(\phi)}\eta_{\mu\nu}dx^\mu dx^\nu + r_c^2 d\phi^2, \tag{25}$$

where $\phi = 0$ and π is the boundary of the fifth dimension. The gravity action in the bulk is expressed as

$$S_{gravity} = \int d^4x \int_{-\pi}^{\pi} d\phi \sqrt{-G} -\Lambda + 2M^3 R, \tag{26}$$

when the $\sigma(\phi)$ is expressed as

$$\sigma(\phi) = r_c|\phi|\sqrt{\frac{-\Lambda}{24M^3}}, \tag{27}$$

provided appropriate fine tuning of the boundary actions.

The geometry allows us to control the masses of SM particles. If the Higgs boson is at $\phi = \pi$ boundary (which is called visible brane), the kinetic term is expressed as

$$S_{vis} = \int d^4x \sqrt{-\bar{g}_{vis}} e^{-4kr_c\pi} \times \left\{ g_{vis}^{\mu\nu} e^{2kr_c\pi} D_\mu H^\dagger D_\nu H - \lambda(|H|^2 - v_0^2)^2 \right\}. \tag{28}$$

The mass term receives the suppression factor of $e^{-kr_c\pi}$ after rescaling the Higgs field so that they have canonical kinetic terms. By adjusting parameters one can easily obtain the mass of the SM particle of the oder of the EW scale while all parameters of the fundamental fifth dimensional Lagrangian are of the order of M_{pl} without fine turning.

The model predicts towers of KK particles with mass of the order of $\Lambda_\phi = \sqrt{6}M_{pl}e^{-kr_c\pi}$ for the particles living in the fifth dimension(bulk). A popular set up of the model is that all the SM fermions are the zero mode of the particles living in the bulk, and the Higgs boson lives in the IR brane. Mass term of the fifth dimensional Lagrangian of the SM model matters control the profile of the fields in the bulk. One can adjust the mass so that light (heavy) quarks and lepton have small (large) overlap with the IR brane so that Yukawa couplings in the four dimensional effective Lagrangian is realized without introducing too much hierarchy among the interactions between the Higgs boson and the bulk fermions. There are on-going search of the KK gauge bosons and KK fermions at the LHC, however, FCNC constraints require $\Lambda_\phi > 10$ TeV already, and it is unlikely that these new particles will be found at the LHC.

7 Suggested reading

To those who is interested in Supersymmetry, a good review for start with is S. P. Martin,"A Supersymmetry primer," In *Kane, G.L. (ed.): Perspectives on supersymmetry II* 1-153 [hep-ph/9709356]. For A review of composite Higgs model, I suggest R. Contino, "The Higgs as a Composite Nambu-Goldstone Boson," arXiv:1005.4269 [hep-ph].

Flavour Physics and CP Violation

Emi Kou
Laboratoire de l'Accélérateur Linéaire, Université Paris-Sud 11, CNRS/IN2P3, Orsay, France

Abstract
In these three lectures, I overview the theoretical framework of the flavour physics and CP violation. The first lecture is the introduction to the flavour physics. Namely, I give theoretical basics of the weak interaction. I follow also some historical aspect, discovery of the CP violation, phenomenological studies of charged and neutral currents and the success of the GIM mechanism. In the second lecture, I describe the flavour physics and CP violating phenomena in the Standard Model (SM). I also give the latest experimental observation of the CP Violation at the B factories and the LHC and discuss its interpretation. In the third lecture, I discuss the on-going search of the signals beyond SM in the flavour physics and also the future prospects.

1 Introduction

The Standard Model (SM) is a very concise model and at the same time a very successful chapter in particle physics. In the establishment of the SM, the flavour changing and/or the CP violating phenomena had played a crucial roles. On the other hand, there is a very important unsolved question related to the CP violation: *how the matter and anti-matter asymmetry of the universe occurs in the evolution of the universe?* Although the Kobayashi-Maskawa mechanism has been successful to explain the CP violation in the flavour phenomena, it is known that the single complex phase introduced in this mechanism is *not* enough to solve this problem. Since there is no known way to introduce another source of CP violation in the SM (except for the strong CP phase), we expect that the SM needs to be extended. Apart from this issue, there are various reasons to expect physics beyond the SM. The search for a signal beyond SM is a most important task of particle physics today.

In this lecture, we expose the theoretical basis of flavour physics in the SM and its phenomenology.

2 Weak interaction: fundamentals of flavour physics

2.1 Quarks and leptons

The flavour physics concerns the interaction among different fermions, quarks and leptons. Fermions are known to appear in three generations:

Quarks			
Charge	Generation		
	I	II	III
+2/3e	u up	c charm	t top
-1/3e	d down	s strange	b bottom

Leptons			
Charge	Generation		
	I	II	III
0	ν_e electron neutrino	ν_μ muon neutrino	ν_τ tau neutrino
-e	e electron	μ muon	τ tau

As we will see in the following, the interactions between the fermions with difference of charge ± 1 can be described by the *charged current* while the interactions between the fermions with the same charge is described by the *neutral current*. The examples of such processes are β decays, $K - \bar{K}$ mixing, $e^+\nu$ scattering process, etc... All these processes are governed by an effective coupling, the so-called Fermi constant $G_F = 1.16639(2) \times 10^{-5}$ GeV^{-2}.

2.2 Charged current

The history of the weak interaction started from the observation of the continuum spectrum of the β decay of nucleons in the 1930's:

$$_ZX \rightarrow_{Z\pm1} X + e^{\mp}\nu \tag{1}$$

where ν is the neutrino postulated by Pauli. During the next two decades, many new experiments were performed and new particles and new decays were discovered. In particular, the two particles called θ and τ[1] were quite puzzling. They both contain *strangeness* and have very similar properties. Besides, they have different decay patterns: θ decays into two pions and τ into three pions. For a solution to this problem, Lee and Yang had proposed the *parity violation* of the weak interaction that was successfully tested by Wu through the β decay of ^{60}Co. After various experimental tests and theoretical argument, it was suggested that the weak interaction should be of the form of $V - A$ (V: Vector current, A: Axial vector current). In this theory, the charged current involves fermions with only left-handed chirality. Thus, the weak interaction processes in which charge is exchanged between leptons and leptons/hadrons are well described at low energy by the effective Lagrangian:

$$\frac{G_F}{\sqrt{2}} J_\mu J^\mu \tag{2}$$

where $J_\mu = \overline{e}\gamma_\mu(1-\gamma_5)\nu$, $J_\mu = \overline{q_d}\gamma_\mu(1-\gamma_5)q_u$. where q_u, q_d are the up and down type quarks, respectively. One of the problems of this theory at the early time was that the discrepancy in the vector coupling when measuring the decay of radioactive oxygen, ^{14}O: the coupling constant which was thought to be universal, G_F, which is the case for the lepton current, was $0.97G_F$. In the early 60's, this problem was nicely understood by introducing the so-called Cabibbo angle θ_c: the coupling of π and K are different and it is proportional to $\cos\theta_c$ and $\sin\theta_c$, respectively. Therefore, the hadronic current (with three quarks) is written as:

$$J_\mu^{\text{hadron}} = \cos\theta_c \overline{u}_L\gamma_\mu d_L + \sin\theta_c \overline{u}_L\gamma_\mu s_L \tag{3}$$

The measurements of $^{14}O \rightarrow^{14} N + e^+\nu$ and $K \rightarrow \pi^0 e^+\nu$ lead to a consistent value of the *Cabibbo angle*, $\theta_c = 0.220 \pm 0.003$, which proved the correctness of this expression.

2.3 Neutral current

From the theory with three quarks, up, down, strange, described above, we can conclude that the quarks provide an $SU(2)$ doublet such as:

$$Q_L = \begin{pmatrix} u \\ d\cos\theta_c + s\sin\theta_c \end{pmatrix}_L \tag{4}$$

Then, the neutral current, namely the term which is induced by $\overline{Q}_L t_3 Q_L$ (t_3 is the SU(2) generator) would induce the term proportional to $\overline{d}_L\gamma_\mu s_L$ and $\overline{s}_L\gamma_\mu d_L$, representing strangeness changing neutral current which were not seen in experiments. This problem was solved by Glashow, Iliopoulos and Miani in 1970, by introducing a hypothetical fourth quark, c. With this extra quark, one can compose another doublet: $\begin{pmatrix} c \\ -d\cos\theta_c + s\sin\theta_c \end{pmatrix}_L$ with which the problematic strangeness changing neutral currents can be cancelled out at the tree level (GIM mechanism). Note however, such flavour changing neutral current can still occur at the loop level if the up quark and the newly introduced charm quark have significantly different masses. Let us see the example of $K - \overline{K}$ mixing. The diagram is given in Fig. 1. This is indeed the strangeness changing ($\Delta S = 2$) neutral current. The amplitude of this process should

[1] Not to be confused with the τ lepton!

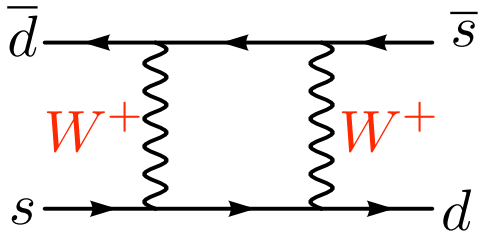


Fig. 1: Feynman diagram inducing $K - \overline{K}$ mixing.

proportional to:

$$G_F^2 \Big[(\sin\theta_c \cos\theta_c)^2 f(m_u) - 2(\sin\theta_c \cos\theta_c)^2 f(m_u, m_c) + (\sin\theta_c \cos\theta_c)^2 f(m_c) \Big] \tag{5}$$

where the first and third terms represent the diagram with either u or c quark in the loop, respectively, while the second term is with both u and c quarks in the loop. The function f is called *loop function*, which contains the result of this loop diagram computation and is a function of the internal particle masses (quarks and W boson in this case). If the mass of the up and charm quarks are the same, the three loop functions in this formula coincide, thus, the full amplitude becomes zero (GIM mechanism at one loop level). In reality, the observed difference in the up and charm quark masses are significantly different, which can yield non-zero $K - \overline{K}$ mixing. What is remarkable about the work by GIM is that the fourth quark, c, was predicted in order to solve the problem of K decays. It took a couple of years since then but indeed the $c\bar{c}$ charm bound state, J/ψ was discovered in 1974.

2.4 Describing the weak interactions in the SM

The $V - A$ theory developed to explain the β decay and strangeness changing interactions is neither renormalizable field theory nor gauge theory. The heavy vector particles which can intermediate the weak interactions is now known as W boson. In the late 60's, the model which unifies the electromagnetic interaction and the weak interaction were developed by S. Glashow, A. Salam and S. Weinberg. In this model, the W, Z and γ can be understood as the gauge bosons of the $SU(2)_L \times U(1)_Y$ gauge group.

In the SM, the masses of the particles are obtained through the Higgs mechanism, where the $SU(2)_L \times U(1)_Y$ symmetry breaks spontaneously to $U(1)_{\rm EM}$ (while keeping the photon massless). Let us see the term which gives the masses of the quarks in the SM, the so-called Yukawa interaction term:

$$\mathcal{L}_Y = \sum_{ij} Y^u_{ij} \overline{\begin{pmatrix} U_i \\ D_i \end{pmatrix}_L} \begin{pmatrix} \phi^0 \\ -\phi^- \end{pmatrix} u_{jR} + \sum_{ij} Y^d_{ij} \overline{\begin{pmatrix} U_i \\ D_i \end{pmatrix}_L} \begin{pmatrix} -\phi^+ \\ \phi^0 \end{pmatrix} d_{jR} + h.c. \tag{6}$$

which is invariant under $SU(3)_C \times SU(2)_L \times U(1)_Y$ gauge transformations. The indices $i, j = 1, 2, 3$ run through the generation. The so-called Yukawa matrix is a completely general complex matrix and not constrained in any way (it could be even non-Hermitian). Then, after the neutral part of the Higgs field acquires the vacuum expectation value, the quark mass matrices are produced:

$$\mathcal{L}_Y = \sum_{ij} m^u_{ij} \overline{U_{iL}} u_{jR} + \sum_{ij} m^d_{ij} \overline{D_{i_L}} d_{jR} + h.c. \tag{7}$$

where

$$m^u_{ij} = Y^u_{ij} \langle \phi^0 \rangle_{\rm vac}, \quad m^d_{ij} = Y^d_{ij} \langle \phi^0 \rangle^*_{\rm vac} \tag{8}$$

This Yukawa mass term can induce parity- and flavour-non-conserving terms. However, we can introduce new quark fields

$$U'_L = K^U_L U_L, \quad u'_R = K^U_R u_R, \quad D'_L = K^D_L D_L, \quad d'_R = K^D_R d_R \tag{9}$$

where the matrices K are constrained only by the condition that they must be unitary in order to preserve the form of the kinetic term. Then, when we re-write the mass term with the prime fields, it takes the same form as above but the new matrix:

$$m^{U\prime} = K_L^U m^U K_R^{U\dagger}, \quad m^{D\prime} = K_L^D m^D K_R^{D\dagger} \tag{10}$$

Now it is a general theorem that for any matrix m, it is always possible to choose unitary matrices A and B such that AmB is real and diagonal. Note here that if the matrix m was Hermitian, we would find $A = B$. Therefore, we choose $m^{U\prime}$ and $m^{D\prime}$ being real and diagonal. Then, the Yukawa mass term does no longer produce the flavour-non-conserving terms, while now the charged current would require some modifications. Let us write the weak doublets with the new prime fields:

$$Q_{iL} = \begin{pmatrix} (K_L^{U-1} U_L')_i \\ (K_L^{D-1} D_L')_i \end{pmatrix} \tag{11}$$

The, the charged current reads:

$$J^{\mu+} = \sum_{i,j} \frac{1}{\sqrt{2}} \overline{U_L^i} \gamma^\mu D_L^j = \sum_{i,j} \frac{1}{\sqrt{2}} \overline{U_L'^i} \gamma^\mu (K_L^{U\dagger} K_L^D)_{ij} D_L'^j = \sum_{i,j} \frac{1}{\sqrt{2}} \overline{U_L'^i} \gamma^\mu V_{ij} D_L'^j \tag{12}$$

where the unitary matrix $V_{ij} \equiv (K_L^{U\dagger} K_L^D)_{ij}$ is known as Cabibbo-Kobayashi-Maskawa matrix. The rotation of the 1-2 part of this matrix corresponds to the Cabibbo angle discussed above. Now it is clear that the quark mixing which differentiates the G_F to $0.97 G_F$ in the hadronic β decay originated from the mismatch between the weak eigenstate and mass eigenstate in the SM.

The full Lagrangian for the quark coupling to the gauge bosons reads[2]:

$$\begin{aligned} \mathcal{L} &= \sum_i \Big[\overline{E_L}(i\partial\!\!\!/) E_L + \overline{l_{iR}}(i\partial\!\!\!/) l_{iR} + \overline{Q_{iL}}(i\partial\!\!\!/) Q_{iL} + \overline{u_{iR}}(i\partial\!\!\!/) u_{iR} + \overline{d_{iR}}(i\partial\!\!\!/) d_{iR} \\ &\quad + g(W_\mu^+ J_W^{\mu+} + W_\mu^- J_W^{\mu-} + Z_\mu^0 J_Z^\mu) + e A_\mu J_{\rm EM}^\mu \Big] \end{aligned} \tag{13}$$

where the coupling g is related to the Fermi constant by $G_F = \frac{g^2}{4\sqrt{2} M_W^2}$. The index $i = 1, 2, 3$ is the generation number. The left handed fermions compose $SU(2)$ doublet as:

$$E_{iL} = \begin{pmatrix} \nu_i \\ L_i \end{pmatrix}_L, \quad Q_{iL} = \begin{pmatrix} U_i \\ D_i \end{pmatrix}_L \tag{14}$$

Note that the assignment of the hypercharge Y is $Y = -1/2$ for E_{iL} and $Y = +1/6$ for Q_{iL}, which together with $T^3 = \pm 1/2$, gives a correct charge $Q = T^3 + Y$. For the right-handed fields, $T^3 = 0$ and thus the hypercharge is equal to the electric charge. Then, the charged, neutral and electro-magnetic currents are written as:

$$J_W^{\mu+} = \frac{1}{\sqrt{2}} (\overline{\nu_{iL}} \gamma^\mu L_{iL} + \overline{U_{iL}} \gamma^\mu D_{iL}) \tag{15}$$

$$J_W^{\mu-} = \frac{1}{\sqrt{2}} (\overline{L_{iL}} \gamma^\mu \nu_{iL} + \overline{D_{iL}} \gamma^\mu U_{iL}) \tag{16}$$

$$\begin{aligned} J_Z^\mu &= \frac{1}{\sqrt{\cos\theta_w}} \Big[\overline{\nu_{iL}} \gamma^\mu (\frac{1}{2}) \nu_{iL} + \overline{L_{iL}} \gamma^\mu (-\frac{1}{2} + \sin^2\theta_w) L_{iL} + \overline{l_{iR}} \gamma^\mu (\sin^2\theta_w) l_{iR} \\ &\quad + \overline{U_{iL}} \gamma^\mu (\frac{1}{2} - \frac{2}{3}\sin^2\theta_w) U_{iL} + \overline{u_{iR}} \gamma^\mu (-\frac{2}{3}\sin^2\theta_w) u_{iR} \end{aligned}$$

[2]Here we show only the result for the first generation but the remaining parts can be derived easily by repeating it with different generations.

$$+\overline{D_{iL}}\gamma^\mu(-\frac{1}{2}+\frac{1}{3}\sin^2\theta_w)D_{iL}+\overline{d_{iR}}\gamma^\mu(\frac{1}{3}\sin^2\theta_w)d_{iR}\,] \quad (17)$$

$$J^\mu_{\rm EM} = \overline{L_i}\gamma^\mu(-1)L_i+\overline{U_i}\gamma^\mu(+\frac{2}{3})U_i+\overline{D_i}\gamma^\mu(-\frac{1}{3})D_i \quad (18)$$

The *weak angle* θ_w relates different couplings and masses, e.g. $g=\frac{e}{\sin\theta_w}$ and $m_W=m_Z\cos\theta_w$.

3 CP violation

3.1 Matter anti-matter asymmetry in nature

Back in 1920's, having the theory of relativity of Einstein, Dirac extended the quantum mechanics to incorporate the matter which moves with close to the speed of light. The relativistic quantum mechanics follows the equation of motion called Dirac equation. This equation had one solution which correspond to the electron and in addition, another one that has the same mass and spin as the electron but with opposite charge, an anti-particle. A couple of years after, in 1932, Anderson discovered a particle in cosmic rays, which indeed corresponds to this solution, a positron! In Dirac's theory, anti-particles and particles can be created and annihilated by pairs. Then, a serious question raises: *why only particles (electron, proton, etc) can exist in the universe but not anti-particles?* This theoretical problem has not been solved yet. It seems that something has happened in the early universe, which caused an unbalance between particles and anti-particles.

Our universe was born about 135×10^{11} years ago, with extremely high temperature, 10^{19} GeV (about 4000 K). After its birth, the universe started expanding. As a result, the temperature dropped rapidly. At the early time when the temperature was high, the high energy photon could pair-create particles and anti-particles (namely, proton/anti-proton, neutron/anti-neutron, electron/anti-electron). At the same time, since all the particles are relativistic, they could also pair-annihilate. As a result, the photon, particle, anti-particle are created and annihilated freely (equilibrium state). Once the temperature reached about 1 MeV, the photon energy was not high enough to create the (anti-)particles. Then, only pair annihilation would have occurred and our universe would not have had any (anti-)particles! However, that has not been the case. For some reasons, by that time, there existed some more particle than anti-particles. The remaining particles composed Helium and then, various nucleus were generated through nuclear interactions. So far, the reason of the asymmetry of number of particle and anti-particle is not known. The only thing we know is that there was some cause of asymmetry when the temperature of the universe was about 10^{15} GeV. And in order for this to happen, there are three conditions (Sakharov's conditions): i) Baryon number violation, ii) *C-symmetry and CP-symmetry violation*, iii) Interactions out of thermal equilibrium.

It turned out that CP symmetry is violated in nature. This is the subject of this section. The observed CP violation in nature is explained well in the framework of the SM. However, it has also been found that the source of CP violation in SM is much too small to explain the matter anti-matter asymmetry of the universe. This is one of the reasons why we strongly believe that there is physics beyond the SM and why we search for further CP violating observables.

3.2 CP violation in the kaon system

The first observation of CP violation was through the measurement of kaon decays. The kaon decays had unusual properties such as the $\theta-\tau$ puzzle as mentioned earlier. The kaons came as two isodoublets (K^+,K^0) and their anti-particles $(K^-,\overline{K}^0)$ with strangeness $+1$ and -1. The difficulty of assigning the θ and the τ to one of K^0 or $\overline{K}^0$ is that θ which decays to two pions should be CP even and τ which decays to three pions should be CP odd while K^0 and $\overline{K}^0$ are both CP even[3]. In 1955, it was proposed

[3]Remember: $\mathcal{CP}|K^0\rangle=|\overline{K}^0\rangle$, $\mathcal{CP}|\overline{K}^0\rangle=|K^0\rangle$, $\mathcal{CP}|\pi^0\rangle=-|\pi^0\rangle$, $\mathcal{CP}|\pi^+\pi^-\rangle=+|\pi^+\pi^-\rangle$, $\mathcal{CP}|(\pi^+\pi^-)_l\pi^0\rangle=(-1)^{l+1}|(\pi^+\pi^-)_l\pi^0\rangle$ where l is angular momentum between $\pi^+\pi^-$ system and π^0.

by Gell-Mann and Pais that the observed state must be the linear combination of the K^0 and $\overline{K}^0$ such as:

$$K_1 = \frac{1}{\sqrt{2}}\left(K^0 + \overline{K}^0\right), \quad K_2 = \frac{1}{\sqrt{2}}\left(K^0 - \overline{K}^0\right). \tag{19}$$

From the weak interaction point of view, this is quite natural since the weak interaction does not distinguish the strangeness, K^0 and $\overline{K}^0$ always mix. Now the problem is solved: K_1 is indeed CP even and K_2 is CP odd[4], which therefore can correspond to θ and τ, respectively. It is important to notice here that the life time of K_1 and K_2 are very different. The masses of K being 498 MeV and π 140 MeV, the three pion final state is suppressed by the small phase (about a factor 600). This reflects in to the lifetime of these particles: $\tau(K_1) \simeq 0.90 \times 10^{-10}\ s$ and $\tau(K_2) \simeq 5.1 \times 10^{-8}\ s$. This *accidental phase space suppression* will play a crucial role for discovering the CP violation in kaon system.

In 1962, the experiment of Cronin, Fitch and his collaborators announced the very surprising result that the long-lived kaon, i.e. K_2, decays into two pions:

$$K_2 \to \pi^+\pi^-$$

Since K_2 is CP odd state while two pion is the CP even state, CP is not conserved in this process! The fraction is rather small, 2×10^{-3} of total charged decay modes. Nevertheless, this is the proof that CP invariance is violated in nature!

A modification to Eq. (19) is in order. Now, we name the short- and long-lived kaons as K_S and K_L, then,

$$K_S = \frac{1}{\sqrt{2}}\left(pK^0 + q\overline{K}^0\right) = \frac{p}{2}\left[(1+\frac{q}{p})K_1 + (1-\frac{q}{p})K_2\right] \tag{20}$$

$$K_L = \frac{1}{\sqrt{2}}\left(pK^0 - q\overline{K}^0\right) = \frac{p}{2}\left[(1-\frac{q}{p})K_1 + (1+\frac{q}{p})K_2\right] \tag{21}$$

CP violation ($K_{S,L} \neq K_{1,2}$) occurs when $q/p \neq 1$.

3.3 Mixing the two kaon states

Let us now formulate these two kaon states in quantum mechanics. The mixing of the two states comes from the weak interaction which changes the flavour. Let us describe the time evolution of the $K\overline{K}$ system in terms of the Hilbert space $|\Psi(t)\rangle = a(t)|K\rangle + b(t)|\overline{K}\rangle$ (here we ignore the multi-particle states). The time dependence of this oscillation can be described by the Schrödinger equation as:

$$i\hbar\frac{\partial}{\partial t}\Psi(t) = \mathcal{H}\Psi(t). \tag{22}$$

where

$$\Psi(t) = \begin{pmatrix} a(t) \\ b(t) \end{pmatrix} \tag{23}$$

The matrix $\mathcal{H}$ is written by

$$\mathcal{H} = \mathbf{M} - \frac{i}{2}\mathbf{\Gamma} = \begin{pmatrix} M_{11} - \frac{i}{2}\Gamma_{11} & M_{12} - \frac{i}{2}\Gamma_{12} \\ M_{21} - \frac{i}{2}\Gamma_{21} & M_{22} - \frac{i}{2}\Gamma_{22} \end{pmatrix} \tag{24}$$

The CPT or CP invariance imposes $M_{11} = M_{22}, \Gamma_{11} = \Gamma_{22}$ and CP or T invariance imposes $\Im\, M_{12} = 0 = \Im\, \Gamma_{12}$. Then, the eigenvalues and the eigenvectors of this matrix read:

$$\text{System 1}: \quad M_{11} - \tfrac{i}{2}\Gamma_{11} + \tfrac{q}{p}\left(M_{12} - \tfrac{i}{2}\Gamma_{12}\right), \quad \begin{pmatrix} p \\ q \end{pmatrix} \tag{25}$$

[4]This choice of the kaon state was originally proposed based on the idea that $\mathcal{C}$ is conserved in weak interaction (notice K_1 and K_2 are $\mathcal{C}$ eigenstates as well). However, when the parity violation in the weak interaction was suggested by Lee and Yang, it was also suggested that charge invariance is also broken, although it was thought that $\mathcal{CP}$ was still a good symmetry in weak interaction.

$$\text{System 2}: \quad M_{11} - \tfrac{i}{2}\Gamma_{11} - \tfrac{q}{p}\left(M_{12} - \tfrac{i}{2}\Gamma_{12}\right), \quad \begin{pmatrix} p \\ -q \end{pmatrix} \tag{26}$$

which leads to

$$|K_1\rangle = p|K\rangle + q|\overline{K}\rangle \tag{27}$$
$$|K_2\rangle = p|K\rangle - q|\overline{K}\rangle \tag{28}$$

with

$$\frac{q}{p} = \pm\sqrt{\frac{M_{12}^* - \frac{i}{2}\Gamma_{12}^*}{M_{12} - \frac{i}{2}\Gamma_{12}}} \tag{29}$$

where the choice of the solution to this equation, either $+$ or $-$, corresponds to replacing the Systems 1 and 2. Now, it became clear that the CP violation $q/p \neq \pm 1$ occurs when M_{12} and/or Γ_{12} is complex number.

The masses and the widths yield:

$$M_1 - \tfrac{i}{2}\Gamma_1 \equiv M_{11} - \tfrac{i}{2}\Gamma_{11} + \tfrac{q}{p}\left(M_{12} - \tfrac{i}{2}\Gamma_{12}\right) \tag{30}$$
$$M_2 - \tfrac{i}{2}\Gamma_2 \equiv M_{11} - \tfrac{i}{2}\Gamma_{11} - \tfrac{q}{p}\left(M_{12} - \tfrac{i}{2}\Gamma_{12}\right) \tag{31}$$

where $M_{1,2}$ and $\Gamma_{1,2}$ are real numbers. Here we choose the $+$ sign for the solution for q/p above, and then we define the mass and the width differences as:

$$\Delta M \equiv M_2 - M_1, \quad \Delta\Gamma = \Gamma_1 - \Gamma_2 \tag{32}$$

These two quantities are very important observables for the mixing system. Note that the discussions are totally general and can apply to $D\overline{D}$ and $B\overline{B}$ systems.

3.4 Time evolution master formula

Now let us describe the time evolution of the kaons decaying into pions. When there is a mixing of two states, these two states oscillate as time evolves. The CP violating phenomena observed in the kaon system implies that the oscillation rate is different for the state which was K at a given time from those with $\overline{K}$. There is another possibility: the CP violation occurs in the decays, i.e. the decay rates of K and $\overline{K}$ are different. To summarize, there are a two possibilities of source of the CP violation:

$$\text{Oscillation}: \; K \overset{\not{C}\not{P}}{\longleftrightarrow} \overline{K}, \;\; \text{and/or} \qquad \text{Decay}: \; (K, \overline{K}) \overset{\not{C}\not{P}}{\longrightarrow} \pi\pi(\pi) \tag{33}$$

Therefore, we are going to derive the time evolution formulae which describe the oscillation and the decays. The oscillation part is already done. It is the solution to the Schrödinger equation given above. The states at time t, starting as K and $\overline{K}$ at $t = 0$ are given:

$$|K(t)\rangle = f_+(t)|K\rangle + \frac{q}{p} f_-(t)|\overline{K}\rangle \tag{34}$$
$$|\overline{K}(t)\rangle = f_+(t)|\overline{K}\rangle + \frac{p}{q} f_-(t)|K\rangle \tag{35}$$

where

$$f_\pm = \frac{1}{2} e^{-iM_1 t} e^{-\frac{1}{2}\Gamma_1 t}\left[1 \pm e^{-i\Delta M t} e^{\frac{1}{2}\Delta\Gamma t}\right] \tag{36}$$

Now the decay part. The decay amplitude of $K/\overline{K}$ to given final state f ($f = \pi\pi$ or $\pi\pi\pi$) can be expressed by the matrix element with effective Hamiltonian with $\Delta S = 1$:

$$A(f) = \langle f|\mathcal{H}_{\Delta S=1}|K^0\rangle, \quad \overline{A}(f) = \langle f|\mathcal{H}_{\Delta S=1}|\overline{K}^0\rangle \tag{37}$$

Then, the decay width of the state which was K^0 and $\overline{K}^0$ at $t = 0$ reads:

$$\begin{aligned}
\Gamma(K^0(t) \to f) &\propto e^{-\Gamma_1 t}|A(f)|^2 \left[K_+(t) + K_-(t)\left|\frac{q}{p}\right|^2 |\overline{\rho}(f)|^2 + 2\Re\left[L^*(t)\left(\frac{q}{p}\right)\overline{\rho}(f)\right]\right] \\
\Gamma(\overline{K}^0(t) \to f) &\propto e^{-\Gamma_1 t}|\overline{A}(f)|^2 \left[K_+(t) + K_-(t)\left|\frac{p}{q}\right|^2 |\rho(f)|^2 + 2\Re\left[L^*(t)\left(\frac{q}{p}\right)\rho(f)\right]\right]
\end{aligned}$$

where

$$\overline{\rho}(f) \equiv \frac{\overline{A}(f)}{A(f)} \equiv \frac{1}{\rho(f)} \tag{38}$$

$$|f_\pm(t)|^2 = \frac{1}{4}e^{-\Gamma_1 t}K_\pm(t) \tag{39}$$

$$f_-(t)f_+^*(t) = \frac{1}{4}e^{-\Gamma_1 t}L^*(t) \tag{40}$$

$$K_\pm(t) = 1 + e^{\Delta\Gamma} \pm 2e^{\frac{1}{2}\Delta\Gamma t}\cos\Delta M t \tag{41}$$

$$L^*(t) = 1 - e^{\Delta\Gamma} + 2ie^{\frac{1}{2}\Delta\Gamma t}\sin\Delta M t \tag{42}$$

The CP violation manifests itself as:

$$\mathcal{A} = \frac{\Gamma(\overline{K}^0(t) \to f) - \Gamma(K^0(t) \to f)}{\Gamma(\overline{K}^0(t) \to f) + \Gamma(K^0(t) \to f)} \neq 0 \tag{43}$$

3.5 The three types of CP violation

In this section, we learn the three types of CP violating processes:

- Direct CP violation (no-oscillation)
- Flavour specific mixing CP violation
- Flavour non-specific mixing CP violation (time dependent CP violation)

Direct CP violation (no-oscillation):

No-oscillation means $\Delta M = 0, \Delta\Gamma = 0$ then, we have $K_-(t) = L(t) = 0$. In this type, CP violation occurs only through the decay:

$$|A(f)| \neq |\overline{A}(\overline{f})| \tag{44}$$

The CP asymmetry is given as:

$$\mathcal{A} = \frac{|\overline{A}(\overline{f})|^2 - |A(f)|^2}{|\overline{A}(\overline{f})|^2 + |A(f)|^2} = \frac{|\overline{\rho}(\overline{f})|^2 - 1}{|\overline{\rho}(\overline{f})|^2 + 1} \tag{45}$$

It should be noted that non-zero CP asymmetry $\mathcal{A} \neq 0$ occurs only when $|\overline{\rho}| \neq 1$ ($\arg(\overline{\rho}) \neq 0$ is not sufficient!).

Flavour specific mixing CP violation :

Let's consider the semi-leptonic decay, e.g. $K^0 \to Xl^+\nu$ or $\overline{K}^0 \to Xl^-\overline{\nu}$. Note that at the level of quark and leptons, these decays come from $\overline{s} \to \overline{u}W^+(\to l^+\nu)$ and $s \to uW^-(\to l^-\overline{\nu})$, respectively. In such a decay mode, the initial state and the final state have one to one correspondence: tagging of the final state flavour (or lepton charge) tells whether the initial state was K^0 or $\overline{K}^0$. Defining the decay amplitude as:

$$A_{SL} \equiv |A(Xl^+\nu)| = |\overline{A}(Xl^-\overline{\nu})| \tag{46}$$

(note i) this equality comes from CPT invariance and ii) $|A(Xl^-\overline{\nu})| = |\overline{A}(Xl^+\nu)| = 0$), we find the decay rates for the state which was K^0 or $\overline{K}^0$ at $t = 0$ read:

$$\Gamma(K^0(t) \to l^+X) \propto e^{-\Gamma_1 t}K_+(t)|A_{SL}|^2 \tag{47}$$

$$\Gamma(K^0(t) \to l^-X) \propto e^{-\Gamma_1 t}K_-(t)\left|\frac{q}{p}\right|^2|A_{SL}|^2 \tag{48}$$

$$\Gamma(\overline{K}^0(t) \to l^-X) \propto e^{-\Gamma_1 t}K_+(t)|A_{SL}|^2 \tag{49}$$

$$\Gamma(\overline{K}^0(t) \to l^+X) \propto e^{-\Gamma_1 t}K_-(t)\left|\frac{p}{q}\right|^2|A_{SL}|^2 \tag{50}$$

where the wrong sign processes (the second and the fourth lines) come from the $K^0 \leftrightarrow \overline{K}^0$ oscillation. The CP asymmetry is given as:

$$\mathcal{A} = \frac{|p/q|^2 - |q/p|^2}{|p/q|^2 + |q/p|^2} = \frac{|p/q|^4 - 1}{|p/q|^4 + 1} \tag{51}$$

which does not depend on the time.

Flavour non-specific mixing CP violation (time dependent CP violation) :
For this type of CP violation to be measured, we utilize very special kinds of final state: the final state to which both K^0 and $\overline{K}^0$ can decay. The CP eigenstate $CP|f_\pm\rangle = \pm|f_\pm\rangle$ falls into this category. Indeed, the $\pi\pi$ final states are such a case:

$$K^0 \to \pi^+\pi^-, \overline{K}^0 \to \pi^+\pi^-, \quad K^0 \to \pi^0\pi^0, \overline{K}^0 \to \pi^0\pi^0 \tag{52}$$

In general, both $|\overline{\rho}(f)| \neq 1$ and $q/p \neq 1$ can occur. Just for simplicity, we present the result for $|\rho(f)| = 1$ and $|q/p| = 1$,

$$\mathcal{A} = \frac{2\sin(\arg q/p + \arg\overline{\rho})e^{\frac{1}{2}\Delta\Gamma t}\sin\Delta Mt}{1 + e^{\Delta\Gamma t} + \cos(\arg q/p + \arg\overline{\rho})[1 - e^{\Delta\Gamma t}]} \tag{53}$$

Thus, the non-zero CP asymmetry will occur when $\arg q/p + \arg\overline{\rho} =\neq$ and $\Delta M \neq 0$. The asymmetry depends on the time in this case. We will come back to this type of CP asymmetry later on the B meson system.

3.6 CP violation in $B\overline{B}$ system

The discovery of the CP violation in K system is helped by the (accidental) fact that the two (supposed-to-be) eigenstates K_S and K_L have very different life time, which allowed us to realize that K_L (CP-odd state) decayed to $\pi\pi$ (CP even state). In the B meson system, of two B states both have very short life time. Thus, we need some strategy to identify whether the initial was B or $\overline{B}$. The most common way to achieve this task is the following:

- $t = 0$: B and $\overline{B}$ are pair-produced from e^+e^- collision (in this way, the $B\overline{B}$ is produced in a C odd configuration).
- $t = t_1$: one of B or $\overline{B}$ decay semi-leptonically. As presented in the previous section, if the final state contained $l^{-(+)}$, then, the particle that decayed was $\overline{B}(B)$. Due to the quantum-correlation, if $l^{-(+)}$ is detected, the other particle which hasn't decayed yet should be $(B)\overline{B}$.
- $t = t_2$: Then, this remaining particle decays to the CP eigenstate, which is common for B and $\overline{B}$. Between $t = t_1$ and $t = t_2$, this particle oscillate between B and $\overline{B}$.

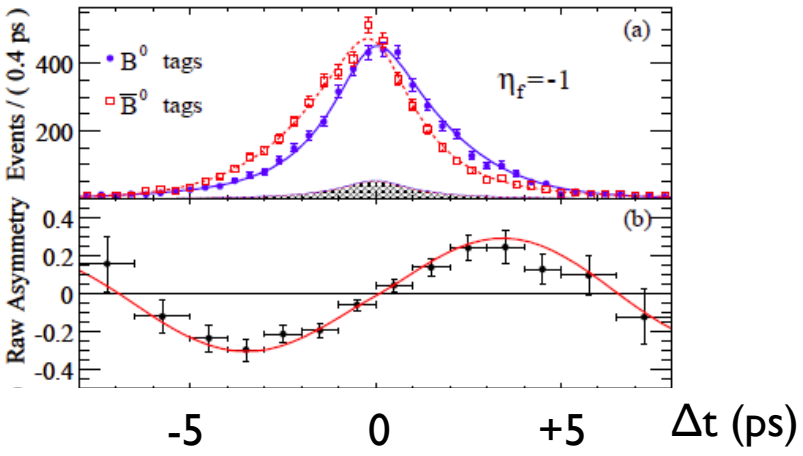


Fig. 2: Time dependence of the $B - \overline{B}$ oscillation.

The decay rate at $t = t_2$ for the processes where we observe $l^{\pm}$ at $t = t_1$ can be written as:

$$\Gamma(B^0(t_2) \to f) \propto e^{-\Gamma_B(t_2-t_1)}|A(B^0 \to f)|^2[1 - \Im(\frac{q}{p}\overline{\rho}(f))\sin(\Delta M_B(t_2 - t_1))]$$
$$\Gamma(\overline{B}^0(t_2) \to f) \propto e^{-\Gamma_B(t_2-t_1)}|A(B^0 \to f)|^2[1 + \Im(\frac{q}{p}\overline{\rho}(f))\sin(\Delta M_B(t_2 - t_1))]$$

where $\overline{\rho} = 1$ is assumed for simplicity and also $\Delta\Gamma_B = 0$ is assumed, which is close to the truth from the observation. If CP is violated $q/p \neq 1$, we should observe different time dependence for these two processes. Indeed, experiment has observed a clear difference between this two and CP violation was confirmed at B factory experiments in 2001 with the final state $f = J/\psi K_S$ (see Fig. 2 top). It was 35 years after the first discovery of CP violation in K decay. In this channel, the time dependent of the asymmetry behaves as:

$$\mathcal{A} = \frac{\Gamma(\overline{B}^0(t) \to f) - \Gamma(B^0(t) \to f)}{\Gamma(\overline{B}^0(t) \to f) + \Gamma(B^0(t) \to f)} = \Im\frac{q}{p}\overline{\rho}(f)\sin\Delta M_B t \tag{54}$$

where $t = t_2 - t_1$. Comparing to the kaon system, the CP violation in B system appeared to be large $\Im\frac{q}{p}\overline{\rho}(J/\psi K_S) \simeq 0.67$.

4 CP violation in SM: unitarity triangle

Now that we have enough evidences of CP violation in nature, both in K and B system. In fact, by now, not only in $K \to \pi\pi(\pi)$ and $B \to J/\psi K_S$ processes, but also CP violation has been observed in many different decay channels. The CP violation for these two channels indicates namely $\arg(q/p) \neq 0$ in K and B system. There are also relatively large *Direct CP violation* observed in various channels (such as $B \to \pi\pi$), which indicates $|\overline{\rho}| \neq 1$ in those channels. A hint of observation of the *flavour-specific CP violation* is also reported ($|q/p| \neq 1$) in B_s system but the experimental result is not precise enough yet. We will discuss on this issue later in this lecture.

In this section, we discuss *where the complex phase comes from* in the SM in order to have $\arg(q/p) \neq 0$. In the model building point of view, it is not easy to incorporate a complex parameter to the theory, namely because, the CP violation is observed only in the K and B systems but nowhere else. Most strong constraint for introducing complex phase to the theory comes from the non-observation

of the electric-dipole moment (EDM) of leptons and neutrons. As we have discussed before, the Yukawa matrix contains free parameters of SM and can be no-Hearmitian. The observable of the Yukawa matrix, the CKM matrix, is only constrained to be unitary, thus can contain a complex number. The CKM matrix is the coupling for the flavour changing charged current, thus, it is ideal to generate CP violation only in the flavour non-diagonal sectors.

4.1 Kobayashi-Maskawa ansatz

The fact that the CKM matrix contains complex phases does not necessarily mean that they generate observable CP violation, since some phases can be absorbed by the redefinition of the field. In 1973, Kobayashi and Maskawa investigated this question. A general $n \times n$ unitary matrix contains $2n^2 - (n + (n^2 - n)) = n^2$ real parameters. The phases of the quark fields can be rotated freely, $\psi \to e^{i\phi}\psi$ (applying separately for up-type and down-type quarks), while one overall phase is irrelevant. Thus, we can rotate $2n - 1$ phases by this. As a result, we are left with $n^2 - (2n - 1) = (n - 1)^2$ real parameters. Among these parameters, we subtract the number of the rotation angles, which is the number of the real parameter in $n \times n$ orthogonal matrix $\frac{1}{2}n(n - 1)$. As a result, the number of the independent phase in CKM matrix is: $\frac{1}{2}(n - 1)(n - 2)$. Kobayashi and Maskawa concluded that *in order for CP to be broken through CKM matrix, third generation of quarks is necessary*. In 1973 when they wrote this paper, there were only three quarks confirmed (up, down and strange) with a speculation of the fourth quark (charm). The prediction of further two quarks was rather bold. However, indeed, the J/ψ (a charm anti-charm bound state) was discovered in 1974. The third generation lepton τ was seen in 1975 and confirmed in 1977. Also in 1977, the fifth quark, bottom was discovered. For the sixth quark, top, are needed to wait until 1994. Now the Kobayashi and Maskawa mechanism is a part of the SM. As we see in the following, all the observed CP violations can be explained by the single phase in the CKM matrix *at a certain level*. Therefore, it is believed that this phase is the dominant source of the observed CP violation.

4.2 The unitarity triangle

As we have repeated, the CKM matrix is restricted by theory only to be unitary. It contains four free parameters (three rotation angles and one phase), which should explain all observed flavour changing and non-changing phenomena including CP violating ones. Thus, the test of the unitarity of the CKM matrix is a very important task in particle physics. For this purpose, let us first inroduce the most commonly used parameterization.

First we write the 3×3 unitary matrix as product of three rotations ordered as:

$$\mathbf{V} = \omega(\theta_{23}, 0)\omega(\theta_{13}, -\delta)\omega(\theta_{12}, 0) \tag{55}$$

$$= \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix} \tag{56}$$

Now we re-define the parameters (Wolfenstein's parameterization):

$$\sin\theta_{12} = \lambda, \quad \sin\theta_{13} = A\sqrt{\rho^2 + \eta^2}\lambda^3, \quad \sin\theta_{23} = A\lambda^2 \tag{57}$$

Then, realizing that the observed CKM matrix elements follow some hierarchy, we expand in terms of $\lambda(\simeq 0.22)$:

$$\mathbf{V} = \begin{pmatrix} 1 - \frac{1}{2}\lambda^2 & \lambda & A\sqrt{\rho^2 + \eta^2}e^{-i\delta}\lambda^3 \\ -\lambda & 1 - \frac{1}{2}\lambda^2 & A\lambda^2 \\ A(1 - \sqrt{\rho^2 + \eta^2}e^{i\delta})\lambda^3 & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}(\lambda^4) \tag{58}$$

In testing whether all the observed FCNC and CP violating phenomena can be explained by the CKM matrix (unitary matrix which can be written in terms of three rotation angles and one complex

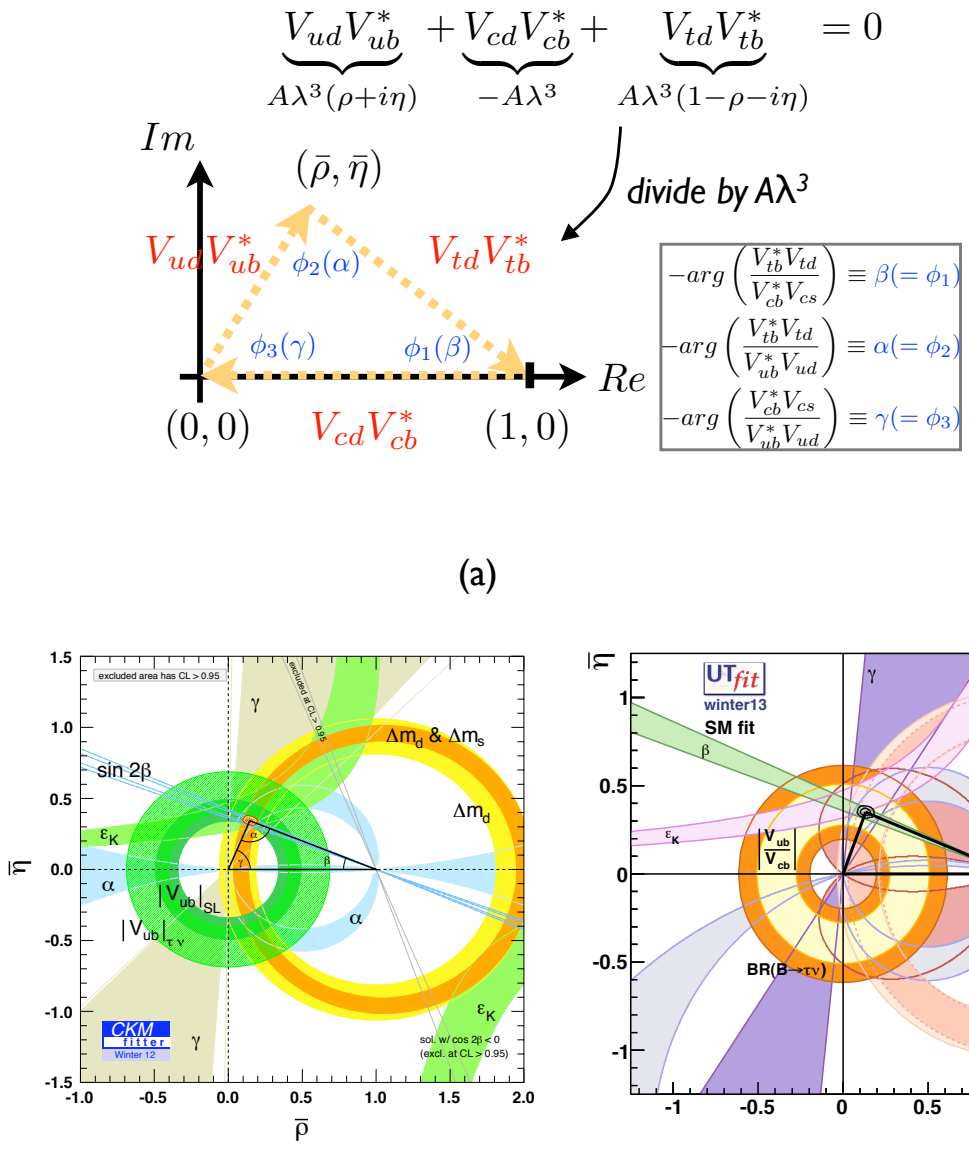


Fig. 3: (a) The unitarity triangle. (b) and (c) The current situation of the unitarity triangle constraints from various flavour observables.

phase), the so-called unitarity triangle is often useful. The unitarity triangle represents one of nine unitarity conditions, the one that contains the matrix elements relevant to B physics:

$$V_{ud}^*V_{ub} + V_{cd}^*V_{cb} + V_{td}^*V_{tb} = 0 \tag{59}$$

Assuming that each term is a vector in the complex plane (known as the $\bar{\rho} = \bar{\eta}$ plane), we can draw a triangle (see Fig. 3 a). We measure *independently* the sides and the angles of this triangle to test, in particular, whether the triangle closes or not. The latest result is shown in Fig. 3 b and c. Let us first look at the measurements of two sides, $|V_{ub}|$ (left side) and $\Delta M_d/\Delta M_s$ (right side). The overlap region of theses two measurements determine roughly the position of the apex of the triangle. One can see that the triangle is *not flat* from these constraints. The one of the three angles, $\beta(=\phi_1)$ is measured very precisely at the B factories through the observation of the B_d oscillation. And this angle is measured

as $(21.7 \pm 0.64)^\circ$, which is indeed rather large. The right side of the triangle drawn by using this value of $\beta(=\phi_1)$, the allowed bound passes through the allowed range from the $|V_{ub}|$ and $\Delta M_d/\Delta M_s$ measurements. Moreover, the overlapping region from these three measurements has also an overlap with the allowed range from the K oscillation measurement, ϵ_K.

The apex of the triangle determined from various measurements is constrained to be in *a small region*, indicating that these phenomena can be explained by the four free parameters of the SM which are in the CKM matrix. In particular, the success of the KM mechanism is manifested by the CP violation in the K and the B systems being explained by the single complex phase in the CKM matrix.

However, the Fig. 3 b and c apparently show that the whole program of verifying the unitarity of the CKM matrix has not been finished yet. The remaining two angles, $\alpha(=\phi_2)$ and $\gamma(=\phi_3)$, have not been measured as precisely as $\beta(=\phi_1)$. Indeed, experimentally, the LHCb experiment has an ability to determine $\gamma(=\phi_3)$ through e.g. $B \to D^{(*)}K$ modes at a higher precision. It will be interesting to see if the right side drawn by using a more precise $\gamma(=\phi_3)$ measurement in the future will still pass through the apex regions allowed by the other measurements. We should like to draw the attention to a *subtle tension* appearing in the Fig. 3 b and c: the overlap region among $|V_{ub}|$, $\Delta M_d/\Delta M_s$ and $\beta(=\phi_1)$. For now, these three bounds have an overlapping region as discussed above. However, the latest determination of $|V_{ub}|$ from the measurement of the branching ratio of $B \to \tau\nu$ turned out to be slightly higher than the ones determined from the semi-leptonic $b \to ul\nu$ decays. If this tendency remains and the $|V_{ub}|$ value shifts towards a larger value, then, the overlap region with $\beta(=\phi_1)$ could be lost. The super B factory, which are now approved project for B physics, has an ability to measure the $B \to \tau\nu$ branching ratio at a much higher precision. Thus, it will not be too long before the hint of this tension will be revealed. Finally, we should also mentioned that the errors indicated in the Fig. 3 b and c contain not only the experimental ones but also the theoretical ones, namely coming from the hadronic uncertainties. And in particular for $|V_{ub}|$ and $\Delta M_d/\Delta M_s$, the theoretical uncertainties are the dominant sources of the error. Thus, in order to achieve a high prevision in determining these parameters, a reduction of the theoretical uncertainty is the most essential. Progresses in various theoretical methods based on QCD, in particular, Lattice QCD, are key for this goal.

5 CP violation in the B_s system: search for physics beyond the SM

5.1 The B_s oscillation

We can derive the B_s oscillation formulae in the same way as B_d system. Experimentally, the following quantities are measured:

$$\Delta M_s \equiv M_2 - M_1 = -2|M_{12}|, \quad \Delta\Gamma_s \equiv \Gamma_1 - \Gamma_2 = 2|\Gamma_{12}|\cos\zeta_s \tag{60}$$

$$\frac{q}{p} \simeq e^{-i\phi_s}\left(1 + \frac{\Delta\Gamma_s}{2\Delta M_s}\tan\zeta_s\right), \quad \left|\frac{q}{p}\right| \simeq 1 + \frac{\Delta\Gamma_s}{2\Delta M_s}\tan\zeta_s \tag{61}$$

where the phases are defined as

$$\phi_s \equiv \arg[M_{12}], \quad \zeta_s \equiv \arg[\Gamma_{12}] - \arg[M_{12}] \tag{62}$$

ΔM_s is measured by Tevatron rather precisely, $\Delta M_s = (17.77 \pm 0.19 \pm 0.07)\ \mathrm{ps}^{-1}$. Recently, many progresses have been made for determining the CP violating phase ϕ_s. In SM, this phase is related to $\phi_s = \beta_s \equiv -arg[V_{tb^*}V_{ts}/V_{cb^*}V_{cs}]$. The SM phase is known to be very small $\beta_s \simeq 2^\circ$, while the LHCb experiment has an ability to reach to this level measuring the B_s oscillation with the $b \to c\bar{c}s$ decay channels such as $B_s \to J/\psi\phi$ or $B \to J/\psi f_0$. It should be noted that different from the B_d system, $\Delta\Gamma_s/\Delta M_s$ is non-negligible. Thus, the precise determination of ϕ_s requires the information of $\Delta\Gamma_s$. Also note that in the case of the $J/\psi\phi$ final state, the angular analysis is required to decompose different polarization state which have different CP. We also have another observable, the B_s oscillation measurement with the lepton final state, namely the di-lepton charge asymmetry A_{sl}, which determines $|q/p|$.

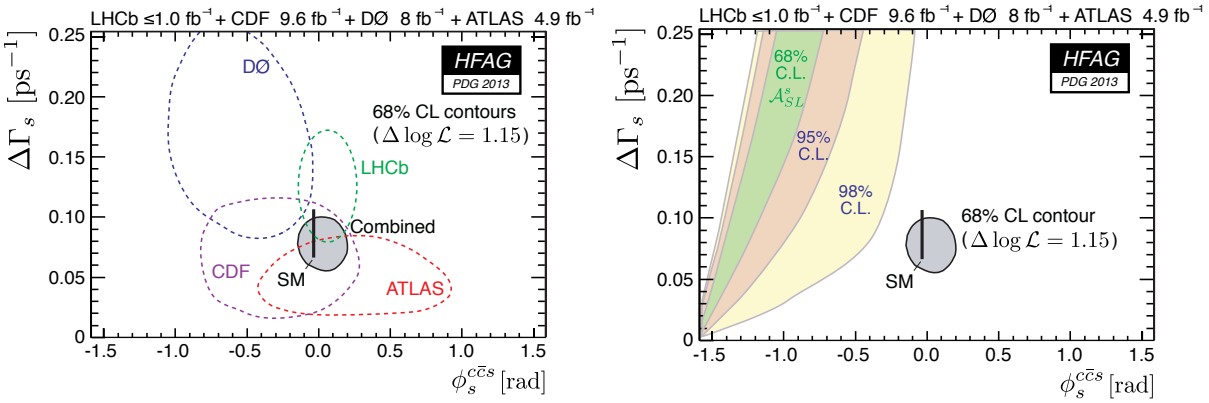


Fig. 4: The current experimental bounds on the B_s oscillation parameters.

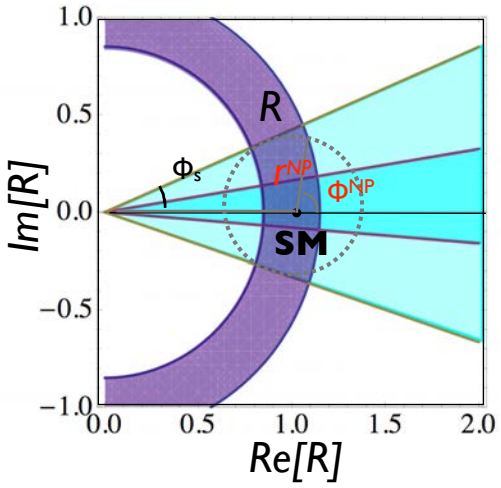


Fig. 5: Illustration of remaining room for new physics after the B_s oscillation phase measurements by LHC (see Eq. 64).

The constraints on ϕ_s and $\Delta\Gamma_s$ from the CDF and the D0 collaboration averaged by HFAG are presented in Fig. 4. On the left figure, the constraint from the B_s oscillation measurement using the $Bs \to J/\psi\phi$ and $Bs \to J/\psi f_0$ final states (contours) and the constraint from the di-lepton charge asymmetry measurement (curved bound) are separately plotted while on the right figure, the combined constraints from these two is presented. It is important to notice that $\Delta\Gamma_s$ is a function of the phase ζ_s as shown in Eq. 61 where ζ_s is related to ϕ_s as in Eq. 62. In particular, as the Γ_{12} is the imaginary part of $B_s - \bar{B}_s$ box diagram, which comes from the up and the charm quark contributions, it is real, unless a new physics contributed to the imaginary part with a non-zero CP violating complex phase. In the case of $\arg\Gamma_{12} = 0$, we have a physical region is such that the $\Delta\Gamma_s$ always decreases from the SM value when $|\phi_s|$ departs from its SM value $0°$. The average of $B_s \to J/\psi\phi$ and $B_s \to J/\psi f_0$ for ϕ_s measurement is obtained as:

$$\phi_s = 0.04^{+0.10}_{-0.13} \tag{63}$$

which is smaller than the previous Tevatron result (2.3σ deviation was announced) and closer to the SM value.

Next we attempt to discuss the implication of these experimental result from theoretical point of view. In order to clarify how large new physics effects can be still allowed after having the constraints

on the B_s oscillation parameters, ΔM_s and ϕ_s, it is useful to introduce the following prameterization:

$$R \equiv \frac{\langle B_s^0|\mathcal{H}_{\rm eff}^{\rm SM} + \mathcal{H}_{\rm eff}^{\rm NP}|\overline{B}_s^0\rangle}{\langle B_s^0|\mathcal{H}_{\rm eff}^{\rm SM}|\overline{B}_s^0\rangle} \equiv 1 + r^{\rm NP} e^{i\phi^{\rm NP}} \tag{64}$$

where NP indicates the new physics contribution. The Fig. 5 is an illustration of the allowed range of real and imaginary part of R from the ΔM_s and ϕ_s measurements. The purple circle represents the constraints from ΔM_s measurement. The bound includes the experimental error as well as the theoretical one, namely coming from the hadronic parameter. Here, it is illustration purpose only, thus, we assumed that the central experimental value of ΔM_s is equal to the SM value. The blue bounds represent the experimental bound coming from ϕ_s (1σ and 3σ errors). The dotted-circle shows the possible new physics contribution to be added to the SM value $R \simeq 1$. What we can see this result is first, even if the experimental value for ΔM_s is close to the SM value, the CP violating phase is allowed to be very different from the SM value ($\simeq 0$). Now that the LHCb results turned out to be close to the SM point, we can also see that 20 % of new physics contribution can be easily accommodated even taking into account only one sigma error. As mentioned earlier, the LHCb has an ability to measure ϕ_s as small as the SM value ($\simeq -2^\circ$). Thus, there is still a plenty of hope that a new physics effect may appear in these measurements in the future.

6 Motivation to go beyond the SM

The Standard Model (SM) is a very concise model which explains a large number observables with a very few parameters. In particular, the agreement of the electroweak precision data with the SM predictions is quite stunning, which shows the correctness of the unified electroweak interaction with the $SU(2)\times U(1)$ gauge symmetry, including its quantum corrections. The crucial prediction of the SM is the existence of the Higgs boson, which is at the origin of the mass of all particles in the SM. LHC has already seen some hint and the best-fitted mass around 126 GeV is also in quite good agreement of the prediction obtained from the electroweak precision data. Furthermore, the agreement of the flavour physics is also impressive. Basically, the free parameters, three rotation angles and one complex phase in the CKM matrix can explain a large number of different experiments, including flavour changing charged/neutral currents as well as CP violating observables.

Some “hints of physics beyond SM" have been reported from time to time, though so far, none of them is significant enough to declare a discovery of a phenomenon beyond the SM. Then, why do we believe there is something beyond?! Indeed, the SM has a few problems. Let us list a few of them here.

- **Higgs naturalness problem**
  We will see this problem more in details later on but basically this problem is related to the question of why the Higgs boson mass scale is so much lower than the Planck mass scale. The quantum corrections to the Higgs mass depend on a cut-off of the theory. If there is no new physics scale below the Planck scale, then the quantum correction become enormous unless there is an incredible fine-tuning cancellation with the bare mass. But that would be quite unnatural.
- **The origin of the fermion mass hierarchy**
  In the SM there are 19 free parameters: three gauge coupling, one strong CP-violating phase, nine fermion masses, four parameters in the CKM matrix, and two parameters in the Higgs potential. We realize that the Yukawa interaction leads to a large number of these parameters (13 out of 19). Some people find this fact quite unsatisfactory. In particular, these values look quite random although with some hierarchy (e.g. top quark and up or down quark have mass scale difference of order 10^5). A symmetry in the Yukawa interaction has been searched for a while, but there has been so far no obvious solution found.
- **The Strong CP problem**
  Another problem concerns the one of the 19 parameters mentioned above, the strong CP-violating

phase. This phase is experimentally found to be extremely small from bounds on the neutron Electric Dipole Moment (nEDM) while theoretically there is no reason why this should be so. In nature, the observed CP violation effects are all in the flavour non-diagonal sector (such as $K - \overline{K}$ or $B - \overline{B}$ oscillation) while CP violation effects in the flavour diagonal sector (such as the EDM) seems to be extremely small, if not zero. The reason for this has been searched for in relation to the conjectured flavour symmetry mentioned above.

– **The baryon asymmetry of the universe**
It should also be mentioned that it is known that the CP violation is related to another problem, the baryon asymmetry of the universe, the unbalance between matter and anti-matter that occurred in the early universe.

– **Quantum theory of gravity**
Although it is obvious that there is a fourth interaction, the gravitational force, the SM does not incorporate this force. In fact, the quantization of gravity itself is a problem which has no solution yet.

These problems are among the sources of the motivation to go beyond the SM. Theoretically, various types of models are proposed in order to solve one or more of the problems mentioned above. Experimentally also, tremendous efforts are payed to search for a signal beyond the SM.

7 Flavour problems in model building beyond the SM

One can extend the SM by introducing new fields and new interactions. The Lagrangian for these new contributions should follow certain rules (the most fundamental one, e.g. is Lorentz invariance). When adding the new terms, the most important task is to verify that these new terms do not disturb the fantastic agreement of various experimental observations with the SM predictions. If the new physics enters at much higher energy than the SM, then this condition could be naturally satisfied: if the currently observed phenomena are not sensitive to such high scale, the SM is valid as an effective theory.

However, this often means that the new physics scale is extremely high (much beyond the TeV scale which can be reached by the current accelerators) or the couplings between new physics particles and the SM particles are very weak. To set a new physics scale high can be inconvenient for the new physics model building. In particular, for those models which are constructed on the motivation for Higgs naturalness problem, having another large scale much higher than the electroweak scale does not sound very preferable. Therefore, in most of the new physics models, the latter solution, to assume the flavour coupling to be very small, is applied, although it is rather artificial (comparing to the SM where such adjustment was not needed, e.g. to suppress FCNC or to explain the source of CP violation. For example, let us consider that the $K_L - K_S$ mass difference comes from the effective four-Fermi interaction :

$$\frac{g^2}{M^2}\overline{\psi}_i\Gamma_\mu\psi_i\overline{\psi}_j\Gamma^\mu\psi_j \tag{65}$$

If we assume the coupling to be of order 1, we find the new physics scale to be $10^3(10^4)$ TeV (the number in parenthesis corresponds to the case when the so-called chiral-enhancement occurs) while if we assume the coupling is SM-like $g \simeq V_{td}^* V_{ts}$ then, the scale can be down to a few (few hundred) TeV. However, to make a very strong assumption for flavour coupling is not appropriate when we are looking for a new physics signal.

In the following, we see in some details, how the extra flavour violation and CP violation occur in the concrete models and which are the solutions. In general, the new physics models which encounter a serious problem from flavour physics induce tree level flavour changing neutral current (FCNC) or new sources of CP violation.

7.1 Two Higgs Doublet Models

The SM consists of a single Higgs doublet which breaks the electroweak symmetry sector of the SM and gives to the particles their masses from the Yukawa interactions. On the other hand, this feature is retained also if there is more than one Higgs doublet. However, once more than one Higgs is introduced, there are extra sources of CP violation (spontaneous CP violation) and also the extra neutral Higgs can induce Flavour Changing Neutral Current (FCNC). In the following, we briefly review how these extra terms appear and the common solution to suppress them by imposing a discrete symmetry based on the so-called Natural Flavour Conservation.

The Two Higgs Doublet Model (2HDM) is the simplest extension of the standard $SU(2) \times U(1)$ model introducing one more Higgs doublet:

$$\phi_1 = \begin{pmatrix} \phi_1^+ \\ \phi_1^0 \end{pmatrix}, \quad \phi_2 = \begin{pmatrix} \phi_2^+ \\ \phi_2^0 \end{pmatrix} \tag{66}$$

The most general Higg potential for this model can be written as:

$$\begin{aligned} V(\phi_1,\phi_2) &= -\mu_1^2\phi_1^\dagger\phi_1 - \mu_2^2\phi_2^\dagger\phi_2 - (\mu_{12}^2\phi_1^\dagger\phi_2 + h.c.) \\ &\quad +\lambda_1(\phi_1^\dagger\phi_1)^2 + \lambda_2(\phi_2^\dagger\phi_2)^2 + \lambda_3(\phi_1^\dagger\phi_1\phi_2^\dagger\phi_2) + \lambda_4(\phi_1^\dagger\phi_2)(\phi_2^\dagger\phi_1) \\ &\quad \frac{1}{2}\left[\lambda_5(\phi_1^\dagger\phi_2)^2 + h.c.\right] + \left[(\lambda_6\phi_1^\dagger\phi_1 + \lambda_7\phi_2^\dagger\phi_2)(\phi_1^\dagger\phi_2) + h.c.\right] \end{aligned} \tag{67}$$

where the quadratic couplings μ_i have a mass dimension two. After imposing the Hermiticity of the potential, we find that $\mu_{12}, \lambda_{5,6,7}$ can be complex. After the spontaneous symmetry breaking, the Higgs fields obtain the non-zero vacuum expectation values which are invariant under $U(1)$ gauge symmetry:

$$\langle\phi_1\rangle = \begin{pmatrix} 0 \\ v_1 \end{pmatrix}, \quad \langle\phi_2\rangle = \begin{pmatrix} 0 \\ v_2 e^{i\alpha} \end{pmatrix} \tag{68}$$

The two VEV's, $v_{1,2}$ can have each associated phases $\delta_{1,2}$ while since the potential in Eq. 67 depends only on one combination, we can rotate the basis giving one single phase $\alpha \equiv \delta_2 - \delta_1$. Non-zero α induces an extra source of CP violation on top of the complex phase in the CKM matrix. Being $v_{1,2}$ the values where the potential has an stable minimum, the expectation value of the potential

$$\begin{aligned} V_0 &= \mu_1^2v_1^2 + \mu_2^2v_2^2 + 2\mu_{12}^2v_1v_2\cos(\delta_3+\alpha) \\ &\quad +\lambda_1v_1^4 + \lambda_2v_2^4 + (\lambda_3+\lambda_4)v_1^2v_2^2 + 2|\lambda_5|v_1^2v_2^2\cos(\delta_5+2\alpha) \\ &\quad +2|\lambda_6|v_1^3v_2\cos(\delta_6+\alpha) + 2|\lambda_7|v_1v_2^3\cos(\delta_7+\alpha) \end{aligned} \tag{69}$$

should be stable with respect to a variation of α, i.e. $\partial V/\partial\alpha = 0$. Note that δ_i are the complex phases of λ_i. This relation can be used to analyze the condition to have non-zero α. For example, in the case when all the couplings are real, i.e. $\delta_i = 0$, this relation leads to

$$\cos\alpha = -\frac{\lambda_6v_1^2 + \lambda_7v_2^2}{4\lambda_5v_1v_2} \tag{70}$$

Thus, CP can be broken spontaneously without an explicit CP violating phase in the Higgs coupling.

Now, we see the Yukawa coupling of the two Higgs doublet model:

$$\mathcal{L}_Y = \sum_{ij}\overline{\begin{pmatrix} U_i \\ D_i \end{pmatrix}}_L(F_{ij}\tilde{\phi_1} + F'_{ij}\tilde{\phi_2})u_{jR} + \overline{\begin{pmatrix} U_i \\ D_i \end{pmatrix}}_L(G_{ij}\phi_2 + G'_{ij}\phi_1)d_{jR} + h.c. \tag{71}$$

where $\tilde{\phi}_i \equiv i\tau_2\phi_i^{\dagger T} = \begin{pmatrix} \phi_i^{0\dagger} \\ -\phi_i^- \end{pmatrix}$ with τ_2 being the Pauli matrix. After the neutral Higgs acquiring vevs, we find

$$\mathcal{L}_Y = \sum_{ij}\overline{\begin{pmatrix} U_i \\ D_i \end{pmatrix}}_L m_{ij}^u u_{jR} + \overline{\begin{pmatrix} U_i \\ D_i \end{pmatrix}}_L m_{ij}^d d_{jR} + h.c. \tag{72}$$

where

$$m_{ij}^u \equiv F_{ij}\langle\tilde{\phi_1}\rangle + F'_{ij}\langle\tilde{\phi_2}\rangle, \quad m_{ij}^d \equiv G_{ij}\langle\phi_2\rangle + G'_{ij}\langle\phi_1\rangle \tag{73}$$

Now as have done in the SM, we diagonalize the matrix $m_{ij}^{u,d}$ by the inserting unitary matrices $K_{L,R}^{U,D}$ (we need to use different matrices from left and right multiplication unless $m_{ij}^{u,d}$ are hermitian). But here, we see a difference; when we look at the mass basis of Eq. 71 by inserting these matrices K, $F_{ij}^{(\prime)}$ and $G_{ij}^{(\prime)}$ are not necessarily diagonalized simultaneously, and this leads to flavour changing neutral Higgs exchanges. This can induce large FCNC contributions, which could contradict the experimental observations.

One of the most common ways to avoid this problem is to impose the following discrete symmetry:

$$\phi_1 \to -\phi_1, \quad \phi_2 \to \phi_2, \quad d_R \to -d_R, \quad u_R \to u_R \tag{74}$$

which prevents ϕ_1 from coupling to u_R and ϕ_2 to d_R. As a result, the FCNC can be avoided. This model is often called Type II two Higgs doublet model. As the Minimal Supersymmetric Model (MSSM) or the Peccei-Quinn models give the same phenomenological consequences, this models is the most worked type among the two Higgs doublet models. We should also note that imposing the discrete symmetry Eq. 74 to the Higgs potential, the terms proportional to $\lambda_{6,7}$ and μ_{12} are forbidden. Then, from Eq. 70, we find $\alpha = \pm\pi/2$. This solution is equivalent to change the definition $\phi_2 \to i\phi_2$, thus ϕ_1 and ϕ_2 have opposite CP. Nevertheless, it is found that this solution can not lead to an observable CP violation.

Having suppressed the phenomenologically unacceptable CP violation and FCNCs by the discrete symmetry, the main effects in flavour physics are due to the charged Higgs contributions. Even though the LHC searches for the charged Higgs directly, the constraints on the property of this new particle, its mass and its couplings, come mainly indirectly from B decays. The branching ratio measurement of $B \to X_s\gamma$ constrains the mass of charged Higgs to be higher than 295 GeV. Further constraint is expected to be obtained from the branching ratio measurement of $B \to \tau\nu$ (see Fig. 6).

7.2 The (extended) technicolor model

The technicolor model is one of the earliest examples of the dynamical breaking of the electroweak symmetry. The model was constructed in a close relation to QCD. In QCD, the $SU(2)$ chiral symmetry is broken spontaneously at the scale $f_\pi \simeq 93$ MeV, which reflects the fact that at the scale $\Lambda_{\rm QCD}$ the QCD interaction becomes strong. Suppose, then, that there are fermions belonging to a complex representation of a new gauge group, technicolor, $SU(N_{\rm TC})$, whose coupling $\alpha_{\rm TC}$ becomes strong at $\Lambda_{\rm TC}$ around electroweak scale. Then, the relation: $M_W = M_Z\cos\theta_W = \frac{1}{2}gF_\pi$ holds, where $F_\pi \simeq \Lambda_{\rm TC}$ just like in QCD. This model can nicely solve the naturalness problem since the all the produced technihadrons have masses around $\Lambda_{\rm TC}$ and they do not receive a large mass renormalization. However, the model is not complete unless it can provide masses to the SM fermions. For this purpose, an extension of the gauge group was proposed (Extended Technicolor Model (ETM)) that embeds flavour and technicolor into a larger gauge group. The flavour gauge symmetries are broken at a higher scale than the technicolor breaking, $\Lambda_{\rm ETC} \simeq M_{\rm ETC}/g_{\rm ETC}$ where $M_{\rm ETC}$ is the typical flavour gauge boson mass. Then, the generic fermion masses are now given by:

$$m_q(M_{\rm ETC}) \simeq m_l(M_{\rm ETC}) \simeq 2\frac{g_{\rm ETC}^2}{M_{\rm ETC}^2}\langle\overline{T}_L T_R\rangle_{\rm ETC} \tag{75}$$

where T is the technifermion and $\langle\overline{T}_L T_R\rangle_{\rm ETC}$ is the vacuum expectation value. However, an acquisition of the SM fermion mass by coupling to the technifermion can induce a serious flavour problem: the transition $q \to T \to q'$ or $q \to T \to T' \to q'$ produce FCNC. Then, for example, the K mass difference limit and the ϵ_K measurement leads to the mass limit:

$$\frac{M_{\rm ETC}}{Re(\delta_{ds})g_{\rm ETC}} \lesssim 10^3 \text{ TeV}, \quad \frac{M_{\rm ETC}}{Im(\delta_{ds})g_{\rm ETC}} \lesssim 10^4 \text{ TeV}, \tag{76}$$

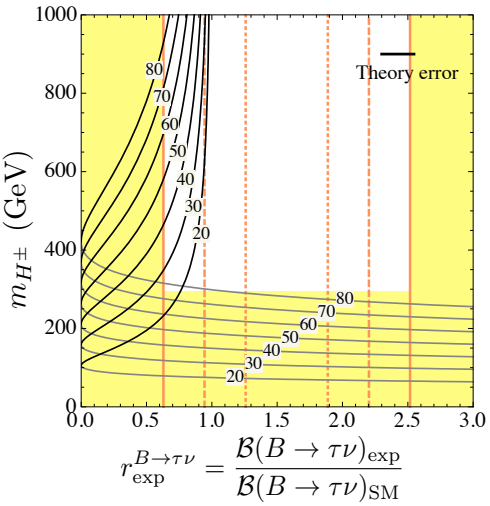


Fig. 6: The charged Higgs contribution to $Br(B \to \tau\nu)$ compared to the present experimental value (normalized to the SM prediction, x-axis) as a function of charged Higgs mass (y-axis). The vertical regions are excluded by the current world average of experimental value, $Br(B \to \tau\nu)_{\rm exp} = (1.15 \pm 0.23) \times 10^{-6}$, at 95% C.L., while the 1σ, 2σ, 3σ errors on the same experimental value are denoted by the grey dotted, dashed, and solid lines, respectively. The horizontal region is excluded by the $Br(B \to X_s\gamma)$ measurement. The grey and the black lines correspond to the two possible solutions, with labels denoting the value of $\tan\beta$. The second solution can lead to a stronger constraint than the one from $B \to X_s\gamma$ especially for large values of $\tan\beta$.

.

respectively. This value together with 75, we find $\Lambda_{\rm TC}$ to be 10-1000 times smaller than the electroweak scale (depending on the flavour coupling δ_{ds}). Several solutions to this problem have been proposed. For example, the so-called "walking technicolor model" induces a large anomalous dimension which enhances the value of fermion masses by keeping the ETC scales relatively low. This can help to reduce the FCNC for the first two generations while the top quark remains problematic, for instance in FCNC processes involving top-quark loops. A possible solution is to generate the small fermion masses by ETC, whereas the top-quark mass is dynamically generated via top condensation (Top-color assisted Technicolor model).

7.3 Supersymmetry

The supersymmetric (SUSY) extensions of the SM are one of the most popular NP models. SUSY relates fermions and bosons. For example, the gauge bosons have their fermion superpartners and fermions have their scalar superpartners. SUSY at the TeV scale is motivated by the fact that it solves the SM hierarchy problem. The quantum corrections to the Higgs mass are quadratically divergent and would drive the Higgs mass to Planck scale $\sim 10^{19}$ GeV, unless the contributions are cancelled. In SUSY models they are cancelled by the virtual corrections from the superpartners. The minimal SUSY extension of the SM is when all the SM fields obtain superpartners but there are no other additional fields. This is the Minimal Supersymmetric Standard Model (MSSM). SUSY cannot be an exact symmetry since in that case superpartners would have the same masses as the SM particles, in clear conflict with observations. If supersymmetry is the symmetry of nature, the masses of the SUSY particles should be the same as their partners'. However, no candidate for SUSY particle has been detected by experiments so far. This indicates that a more realistic model should contain SUSY breaking terms. Different mechanisms of SUSY breaking have very different consequences for flavor observables. In complete generality the MSSM has more than one hundred parameters, most of them coming from the soft SUSY breaking terms

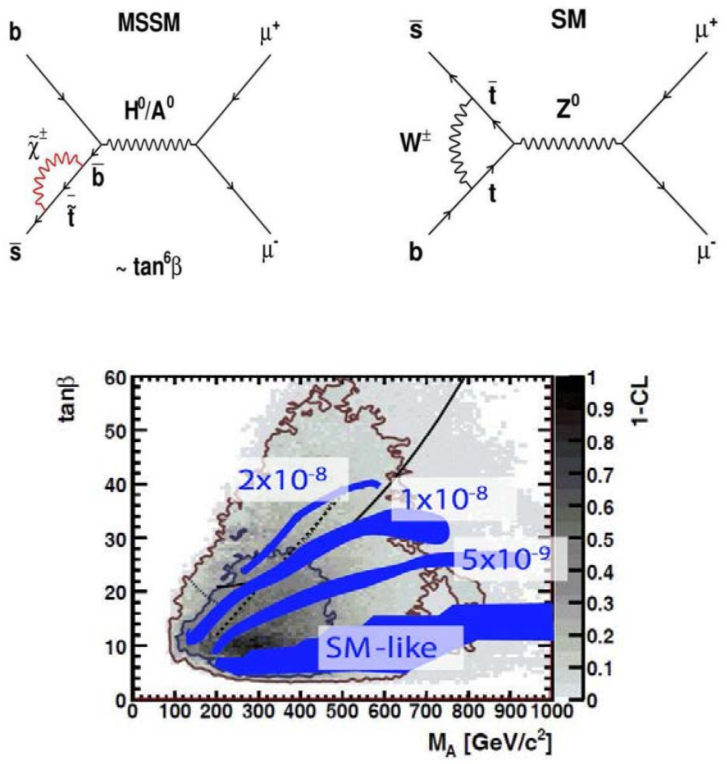


Fig. 7: The Feynman diagram for the $B_{s,d} \to \mu^+\mu^-$ in SM (top right) and in SUSY (top left). The constraint on the SUSY parameter obtained by using the latest LHCb result is also shown (bottom).

.

– the masses and mixings of the superpartners. If superpartners are at the TeV scale the most general form with $\mathcal{O}(1)$ flavor breaking coefficients is excluded due to flavor constraints. This has been called the SUSY flavor problem (or in general the NP flavor problem).

A popular solution to the SUSY flavor problem is to assume that the SUSY breaking mechanism and the induced interactions are flavour "universal". The flavour universality is often imposed at a very high scale corresponding to the SUSY breaking mechanism. It could be at, for instance, the Planck scale ($\sim 10^{19}$ GeV), the GUT scale ($\sim 10^{16}$ GeV) or some intermediate scale such as the gauge mediation scale ($\sim 10^{6}$ GeV). The flavor breaking can then be transferred only from the SM Yukawa couplings to the other interactions through renormalization group running from the higher scale to the weak scale. As a result, the flavor breaking comes entirely from the SM Yukawa couplings (thus, an example of a concrete Minimal Flavour Violation (MFV) NP scenario). Since the soft SUSY breaking terms are flavor-blind, the squark masses are degenerate at the high energy scale. The squark mass splitting occurs only due to quark Yukawa couplings, where only top Yukawa and potentially bottom Yukawa couplings are large. Thus the first two generation squarks remain degenerate to very good approximation, while the third generation squarks are split.

The MFV can be extended by taking into account the large $\tan\beta$ effect. That large $\tan\beta$ scenario leads to a large new physics effects to some B physics observable and can be well tested experimentally. Most recently, the LHCb experiment has made a great progress in this scenario: observation of the $B_s \to \mu^+\mu^-$ process (3.5σ significance). The $B_{s,d} \to \mu^+\mu^-$ comes from the diagram e.g. like in Fig. 7 (top right). It is extremely rare process with branching ratios:

$$Br(B_s \to \mu^+\mu^-) \quad = \quad (3.54 \pm 0.30) \times 10^{-9} \tag{77}$$

$$Br(B_d \to \mu^+\mu^-) \quad = \quad (0.107 \pm 0.01) \times 10^{-9} \tag{78}$$

On the other hand, in the presence of SUSY, there is another contribution like in Fig. 7 (top left), which can be largely enhanced by a large $\tan\beta$ factor:

$$Br(B_{s,d} \to \mu^+\mu^-)_{\rm SUSY} \propto \frac{m_b^2 m_\mu^2 \tan^6\beta}{m_{A_0}^4} \tag{79}$$

Fig. 7 (bottom) illustrate the constraint obtained by using the latest LHCb result:

$$Br(B_s \to \mu^+\mu^-) = (3.2^{+1.5}_{-1.2}) \times 10^{-9} \tag{80}$$

Although the constraint depend on different models, this result excluded most of the scenario with $\tan\beta \gtrsim 50$.

8 Strategies for New Physics searches in flavour physics

The developments of the particle physics today bore the lack of phenomena which cannot be explained in the SM. In flavour physics, many small deviations from SM (say, at the $2-3\sigma$ level) have been reported in the past. However, definitive conclusions for those observation cannot be given so far. Therefore, the strategy in flavour physics is clear: to search for a significant enough deviation from the SM. We tackle this task from two directions, first, to improve the precision of the theoretical prediction, second, to improve the experimental precision. We should emphasize that the latter efforts include not only experimental development, but also to propose theoretically new observables which are sensitive to new physics contributions.

Let us see the example of the CKM unitary triangle Fig. 3. The measurement of the angle, e.g. β, has been improved dramatically the past 10 years since the B factories started. However, the sides measurements (V_{ub}, V_{cb}, V_{tb} etc) have not improved as much since it depends on theoretical inputs and assumptions, namely from the strong interaction. In the future, there are some experimental propositions to improve the experimental measurements. The angle measurements, such as β, γ, will be improved further (γ could be determined as precise as β), can directly be used to improve our knowledge of new physics. The usefulness of the sides measurements relies strongly on progresses in theory, in particular, the effective theory of QCD, lattice QCD or more phenomenological models.

In Tables 1 to 4 we list the new physics sensitive observable. It is again amazing that all these experimental measurements agree within the theoretical and experimental errors so far. On the other hand, the LHCb experiment as well as Belle II experiment will provide us a large samples of new data in coming years.

Table 1: Examples of new physics sensitive observables in B physics. The experimental values are extracted from HFAG (mainly preparation for PDG 2013). The prospects are extracted from [5] for Belle II and [6] for LHCb. The number for Belle II corresponds to the sensitivity at 50 ab^{-1} which can be reached by the year 2023 if the commissioning starts in the year 2015 (physics run in 2016) as scheduled. For LHCb the number corresponds to the sensitivity reach at the year 2018 and the number in parethethesis is for LHCb up-grade.

Observable	Experimental value	Prospect	Comments
$S_{B\to J/\psi K_S} = \sin 2\phi_1(2\beta)$	0.665 ± 0.024	± 0.012 (Belle II) $\pm 0.02(0.007)$ (LHCb)	The current measurement agrees with the SM prediction obtained from the $\phi_1(\beta)$ value extracted using the unitarity relation. A higher precision measurement on $\phi_1(\beta)$ together with the measurements for the other variables in unitarity relation could reveal a new physics contribution. New physics example: bd box diagram and/or tree FCNC
$S_{B\to\phi K_S}$	$0.74^{+0.11}_{-0.13}$	± 0.029 (Belle II) $\pm 0.05(0.02)$ (LHCb)	In the year 2002, a 2-3 σ deviation from the $S_{B\to J/\psi K_S}$ was announced though it is diminished by now. The deviation from $S_{B\to J/\psi K_S}$ is an indication of the CP violation in the decay process of $B \to \phi K_S$, which comes almost purely from the penguin type diagram. New physics example: $b \to s$ penguin loop diagram
$S_{B\to\eta' K_S}$	0.59 ± 0.07	± 0.020 (Belle II)	In the year 2002, a 2-3 σ deviation from the $S_{B\to J/\psi K_S}$ was announced though it is diminished by now. The deviation from $S_{B\to J/\psi K_S}$ is an indication of the CP violation in the decay process of $B \to \eta' K_S$. This decay also comes mainly from the penguin type diagram though this can be only proved by knowing the property of η' (quark content etc). It should also noticed that the branching ratio of this process turned out to be a few times larger than the similar charmless hadronic B decays and this could also been regarded as a hint of new physics. New physics example: $b \to s$ penguin loop diagram. In particular, contributions that can induce $b \to sgg$ (followed by anomaly diagram $gg \to \eta'$) are interesting candidates.

Table 2: Examples of new physics sensitive observables in B physics II (see caption of Table 1).

Observable	Experimental value	Prospect	Comments
$S_{B_s \to J/\psi\phi} = \sin 2\phi_s$	$\phi_s = 0.04^{+0.10}_{-0.13}$	$\pm 0.025(0.008)$ (LHCb)	The phase of B_s mixing is at the order of λ^4 and very small in SM, ~ -0.02. Before the LHCb started, the Tevatron data was showing a $2-3\sigma$ deviation from the SM, which is not diminished. Since the width difference is not negligible in the B_s system (unlike the B_d system), the width difference measurement has to be done simultaneously (width is less sensitive to the new physics though it is not impossible). The LHCb has an ability to reach to the precision as small as this SM value thus, there is still enough room for new physics. New physics example: CP violation in the $b\bar{s}\bar{b}s$ box diagram and/or tree FCNC.
$S_{B_s \to \phi\phi}$	—	$\pm 0.17(0.03)$ (LHCb)	The deviation from $S_{B_s \to J/\psi\phi(f_0)}$ is an indication of the CP violation in the decay process of $B_s \to \phi\phi$, which comes almost purely from the penguin type diagram. The analysis requires a CP state decomposition by studying the angular dependence of the decay. The each component can include different new physics contributions and it is complementary to the $S_{B \to \phi K_S}$ or $S_{B \to \eta' K_S}$ measurements. In addition, the angular analysis also allows us to test the T-odd asymmetry. New physics example: $b \to s$ penguin loop diagram

Table 3: Examples of new physics sensitive observables in B physics III (see caption of Table 1).

Observable	Experimental value	Prospect	Comments
$S_{B\to K_S\pi^0\gamma}$	-0.15 ± 0.20	± 0.02 (Belle II)	This is one of the golden channel for Belle II experiment where the experimental error is expected to be reduced significantly. Non-zero CP violation is the sign of the contamination of the photon polarization which is opposite to the one predicted in SM. The theoretical error is found to be small (less than a few %) though some authors warm a possible large uncertainties to this value. New physics example: right handed current in $b \to s\gamma$ penguin loop diagram
$S_{B_s\to\phi\gamma}$	—	$\pm 0.09(0.02)$ (LHCb)	Non-zero CP violation is the sign of the contamination of the photon polarization which is opposite to the one predicted in SM. The LHCb with its high luminosity could allow us to study this observable and it is complementary to $S_{B\to K_S\pi^0\gamma}$ above. New physics example: right handed current in $b \to s\gamma$ penguin loop diagram
$B \to K^* l^+ l^-$ (low q^2)	—	~ 0.2 (LHCb) in $A_T^2, A_T^{\rm Im}$	The angular distribution carry various information of new physics ($C_{7,9,10}$ and $C'_{7,9,10}$). In particular, the low q^2 region is sensitive to the photon polarization of $b \to s\gamma$.
$B \to K_1\gamma \to (K\pi\pi)\gamma$	—	~ 6 % (LHCb) ~ 18 % (Belle II) in polarization parameter λ_γ	We can obtain the information of the photon polarization of $b \to s\gamma$ through the angular distribution of K_1 decay. Detailed resonance study can improve the sensitivity to the photon polarization.

Table 4: Examples of new physics sensitive observables in B physics IV (see caption of Table 1).

Observable	Experimental value	Prospect	Comments
$\Delta M_{d,s}$	$(0.510 \pm 0.004)_{B_d}\text{ps}^{-1}$ $(17.69 \pm 0.08)_{B_s}$ ps^{-1}	—	The result is consistent to the SM prediction though it depends strongly on the $\|V_{td,ts}\|$. The error is dominated by the theory mainly coming from the $f_{B_{d,s}}$ and B parameters. New physics example: $b\overline{d}bd, b\overline{s}bs$ box diagram and/or tree FCNC
$Br(B \to \tau\nu)$	$(1.15 \pm 0.23) \times 10^{-6}$	$\pm$ 6% (Belle II)	Up to the year 2010, the world average value was 2-3 σ higher than the SM prediction. The SM value depends on $\|V_{ub}\|$ and f_B. New physics example: charged Higgs.
$\mathcal{R} = \frac{B\to D^{(*)}\tau\nu}{B\to D^{(*)}l\nu}$	$(0.440 \pm 0.057 \pm 0.042)_D$ $(0.332 \pm 0.024 \pm 0.018)_{D^*}$	$\pm$ 2.5% for Br of $D^0\tau\nu$ $\pm$ 9.0% for Br of $D^{\pm}\tau\nu$ (Belle II)	In the year 2012, Babar announced 3σ deviation from the SM. New physics example: charged Higgs.
$Br(B \to X_s\gamma)$	$(3.15 \pm 0.23) \times 10^{-4}$	$\pm$ 6 % (Belle II)	Currently SM prediction (at NNLO) is consistent to the experimental value. The error is becoming dominated by the theoretical ones. The result also depends on V_{ts}. New physics example: $b \to s\gamma$ penguin loop
$Br(B_{d,s} \to \mu^+\mu^-)$	$(< 1.0 \times 10^{-9})_{B_d}$ $((3.2^{+1.5}_{-1.2}) \times 10^{-9})_{B_s}$	$(\pm 0.5(0.15) \times 10^{-9})_{B_s}$ (LHCb)	The result is so far consistent to the SM prediction though it depends on the $\|V_{td,ts}\|$. The Minimal Flavour Violation (MFV) hypothesis leads a relation $\frac{Br_{B_s\to\mu^+\mu^-}}{Br_{B_d\to\mu^+\mu^-}} = \frac{\hat{B}_d}{\hat{B}_s}\frac{\tau_{B_s}}{\tau_{B_d}}\frac{\Delta M_s}{\Delta M_d} \simeq 30$. Nevertheless, a large enhancement in $B_d \to \mu^+\mu^-$ is possible in non-MFV. New physics example: large $\tan\beta$ scenarios
$\phi_3 = \gamma$	$(69^{+17}_{-16})^\circ$ (Babar) $(68^{+15}_{-14})^\circ$ (Belle) $(71.1^{+16.6}_{-15.7})^\circ$ (LHCb)	$\pm 1.5^\circ$ (Belle II) $\pm 4^\circ(0.9^\circ)$ (LHCb)	The angle $\phi_3(= \gamma)$ is extracted from the decay modes, DK, D^*K, DK^* etc. These are tree level decays that it can be less affected by the new physics contributions. Therefore, the $\phi_3(= \gamma)$ measurement can be used together with the other purely tree decays, to determine the "SM" unitarity triangle to test the new physics contributions to the other box/penguin loop dominant modes. The precision which can be reached in the future is quite impressive and it will be one of the most important measurements in order to fully understand the unitarity triangle.

Acknowledgements

We would like to thank to the organizers for giving me the opportunity to give a lecture at AEPSHEP 2012. This work was supported in part by the ANR contract "LFV-CPV-LHC" ANR-NT09-508531 and France-Japan corporation of IN2P3/CNRS TYL-LIA.

References

[1] I. I. Bigi and A. I. Sanda, Cambridge Monogr. Part. Phys. Nucl. Phys. Cosmol. **9** (2000) 1.

[2] S. Weinberg, "The Quantum theory of fields. Vol. 1: Foundations," Cambridge, UK: Univ. Pr. (1995) 609 p

[3] S. P. Martin, "A Supersymmetry primer," In *Kane, G.L. (ed.): Perspectives on supersymmetry II* 1-153 [hep-ph/9709356].

[4] Book of Physics of B Factories: Babar and Belle collaborations, in preparation

[5] The Belle II Collaboration, "Belle II Technical Design Report", KEK Report 2010-1 [arXiv:1011.0352]

[6] The LHCb Collaboration, "LHCb reoptimized detector design and performance : Technical Design Report", CERN-LHCC-2003-030, LHCb-TDR-9

Neutrino Physics

Zhi-zhong Xing
Institute of High Energy Physics and Theoretical Physics Center for Science Facilities,
Chinese Academy of Sciences, Beijing, China

Abstract
I give a theoretical overview of some basic properties of massive neutrinos in these lectures. Particular attention is paid to the origin of neutrino masses, the pattern of lepton flavor mixing, the feature of leptonic CP violation and the electromagnetic properties of massive neutrinos. I highlight the TeV seesaw mechanisms as a possible bridge between neutrino physics and collider physics in the era characterized by the Large Hadron Collider.

1 Finite Neutrino Masses

It is well known that the mass of an elementary particle represents its inertial energy when it exists at rest. Hence a massless particle has no way to exist at rest — instead, it must always move at the speed of light. A massive fermion (either lepton or quark) must exist in both left-handed and right-handed states, since the field operators responsible for the non-vanishing mass of a fermion have to be bilinear products of the spinor fields which flip the fermion's handedness or chirality.

The standard model (SM) of electroweak interactions contains three neutrinos $(\nu_e, \nu_\mu, \nu_\tau)$ which are purely left-handed and massless. In the SM the masslessness of the photon is guaranteed by the electromagnetic $U(1)_{\rm Q}$ gauge symmetry. Although the masslessness of three neutrinos corresponds to the lepton number conservation [1], the latter is an accidental symmetry rather than a fundamental symmetry of the SM. Hence many physicists strongly believed that neutrinos should be massive even long before some incontrovertible experimental evidence for massive neutrinos were accumulated. A good reason for this belief is that neutrinos are more natural to be massive than to be massless in some grand unified theories, such as the ${\rm SO}(10)$ theory, which try to unify electromagnetic, weak and strong interactions as well as leptons and quarks.

If neutrinos are massive and their masses are non-degenerate, it will in general be impossible to find a flavor basis in which the coincidence between flavor and mass eigenstates holds both for charged leptons (e, μ, τ) and for neutrinos $(\nu_e, \nu_\mu, \nu_\tau)$. In other words, the phenomenon of flavor mixing is naturally expected to appear between three charged leptons and three massive neutrinos, just like the phenomenon of flavor mixing between three up-type quarks (u, c, t) and three down-type quarks (d, s, b). If there exist irremovable complex phases in the Yukawa interactions, CP violation will naturally appear both in the quark sector and in the lepton sector.

1.1 Some preliminaries

To write out the mass term for three known neutrinos, let us make a minimal extension of the SM by introducing three right-handed neutrinos. Then we totally have six neutrino fields [2]:

$$\nu_{\rm L} = \begin{pmatrix} \nu_{e{\rm L}} \\ \nu_{\mu{\rm L}} \\ \nu_{\tau{\rm L}} \end{pmatrix} \; , \quad N_{\rm R} = \begin{pmatrix} N_{1{\rm R}} \\ N_{2{\rm R}} \\ N_{3{\rm R}} \end{pmatrix} \; , \tag{1}$$

[1] It is actually the $B-L$ symmetry that makes neutrinos exactly massless in the SM, where B = baryon number and L = lepton number. The reason is simply that a neutrino and an antineutrino have different values of $B-L$. Thus the naive argument for massless neutrinos is valid to all orders in perturbation and non-perturbation theories, if $B-L$ is an exact symmetry.

[2] The left- and right-handed components of a fermion field $\psi(x)$ are denoted as $\psi_{\rm L}(x) = P_{\rm L}\psi(x)$ and $\psi_{\rm R}(x) = P_{\rm R}\psi(x)$, respectively, where $P_{\rm L} \equiv (1-\gamma_5)/2$ and $P_{\rm R} \equiv (1+\gamma_5)/2$ are the chiral projection operators. Note, however, that $\nu_{\rm L} = P_{\rm L}\nu_{\rm L}$ and $N_{\rm R} = P_{\rm R}N_{\rm R}$ are in general independent of each other.

where only the left-handed fields take part in the electroweak interactions. The charge-conjugate counterparts of $\nu_{\rm L}$ and $N_{\rm R}$ are defined as

$$(\nu_{\rm L})^c \equiv \mathcal{C}\overline{\nu_{\rm L}}^T \; , \quad (N_{\rm R})^c \equiv \mathcal{C}\overline{N_{\rm R}}^T \; ; \tag{2}$$

and accordingly,

$$\overline{(\nu_{\rm L})^c} = (\nu_{\rm L})^T\mathcal{C} \; , \quad \overline{(N_{\rm R})^c} = (N_{\rm R})^T\mathcal{C} \; , \tag{3}$$

where $\mathcal{C}$ denotes the charge-conjugation matrix and satisfies the conditions

$$\mathcal{C}\gamma_\mu^T\mathcal{C}^{-1} = -\gamma_\mu \; , \quad \mathcal{C}\gamma_5^T\mathcal{C}^{-1} = \gamma_5 \; , \quad \mathcal{C}^{-1} = \mathcal{C}^\dagger = \mathcal{C}^T = -\mathcal{C} \; . \tag{4}$$

It is easy to check that $P_{\rm L}(N_{\rm R})^c = (N_{\rm R})^c$ and $P_{\rm R}(\nu_{\rm L})^c = (\nu_{\rm L})^c$ hold; namely, $(\nu_{\rm L})^c = (\nu^c)_{\rm R}$ and $(N_{\rm R})^c = (N^c)_{\rm L}$ hold. Hence $(\nu_{\rm L})^c$ and $(N_{\rm R})^c$ are right- and left-handed fields, respectively. One may then use the neutrino fields $\nu_{\rm L}$, $N_{\rm R}$ and their charge-conjugate partners to write out the gauge-invariant and Lorentz-invariant neutrino mass terms.

In the SM the weak charged-current interactions of three active neutrinos are given by

$$\mathcal{L}_{\rm cc} = \frac{g}{\sqrt{2}}\overline{(e\;\mu\;\tau)_{\rm L}}\,\gamma^\mu \begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix}_{\rm L} W^-_\mu + {\rm h.c.} \; . \tag{5}$$

Without loss of generality, we choose the basis in which the mass eigenstates of three charged leptons are identified with their flavor eigenstates. If neutrinos have non-zero and non-degenerate masses, their flavor and mass eigenstates are in general not identical in the chosen basis. This mismatch signifies lepton flavor mixing.

1.2 Dirac neutrino masses

A Dirac neutrino is described by a four-component Dirac spinor $\nu = \nu_{\rm L} + N_{\rm R}$, whose left-handed and right-handed components are just $\nu_{\rm L}$ and $N_{\rm R}$. The Dirac neutrino mass term comes from the Yukawa interactions

$$-\mathcal{L}_{\rm Dirac} = \overline{\ell_{\rm L}} Y_\nu \tilde{H} N_{\rm R} + {\rm h.c.} \; , \tag{6}$$

where $\tilde{H} \equiv i\sigma_2 H^*$ with H being the SM Higgs doublet, and $\ell_{\rm L}$ denotes the left-handed lepton doublet. After spontaneous gauge symmetry breaking (i.e., $SU(2)_{\rm L} \times U(1)_{\rm Y} \to U(1)_{\rm Q}$), we obtain

$$-\mathcal{L}'_{\rm Dirac} = \overline{\nu_{\rm L}} M_{\rm D} N_{\rm R} + {\rm h.c.} \; , \tag{7}$$

where $M_{\rm D} = Y_\nu \langle H \rangle$ with $\langle H \rangle \simeq 174$ GeV being the vacuum expectation value of H. This mass matrix can be diagonalized by a bi-unitary transformation: $V^\dagger M_{\rm D} U = \widehat{M}_\nu \equiv {\rm Diag}\{m_1, m_2, m_3\}$ with m_i being the neutrino masses (for $i = 1, 2, 3$). After this diagonalization,

$$-\mathcal{L}'_{\rm Dirac} = \overline{\nu'_{\rm L}} \widehat{M}_\nu N'_{\rm R} + {\rm h.c.} \; , \tag{8}$$

where $\nu'_{\rm L} = V^\dagger \nu_{\rm L}$ and $N'_{\rm R} = U^\dagger N_{\rm R}$. Then the four-component Dirac spinor

$$\nu' = \nu'_{\rm L} + N'_{\rm R} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} \; , \tag{9}$$

which automatically satisfies $P_{\rm L}\nu' = \nu'_{\rm L}$ and $P_{\rm R}\nu' = N'_{\rm R}$, describes the mass eigenstates of three Dirac neutrinos. In other words,

$$-\mathcal{L}'_{\rm Dirac} = \overline{\nu'} \widehat{M}_\nu \nu' = \sum_{i=1}^3 m_i \overline{\nu_i} \nu_i \; . \tag{10}$$

The kinetic term of Dirac neutrinos reads

$$\mathcal{L}_{\rm kinetic} = i\overline{\nu_{\rm L}}\gamma_\mu\partial^\mu\nu_{\rm L} + i\overline{N_{\rm R}}\gamma_\mu\partial^\mu N_{\rm R} = i\overline{\nu'}\gamma_\mu\partial^\mu\nu' = i\sum_{k=1}^{3}\overline{\nu_k}\gamma_\mu\partial^\mu\nu_k \; , \tag{11}$$

where $V^\dagger V = VV^\dagger = \mathbf{1}$ and $U^\dagger U = UU^\dagger = \mathbf{1}$ have been used.

Now we rewrite the weak charged-current interactions of three neutrinos in Eq. (5) in terms of their mass eigenstates $\nu'_{\rm L} = V^\dagger\nu_{\rm L}$ in the chosen basis where the flavor and mass eigenstates of three charged leptons are identical:

$$\mathcal{L}_{\rm cc} = \frac{g}{\sqrt{2}}\overline{(e\;\mu\;\tau)_{\rm L}}\;\gamma^\mu V \begin{pmatrix}\nu_1\\ \nu_2\\ \nu_3\end{pmatrix}_{\rm L} W^-_\mu + {\rm h.c.} \; . \tag{12}$$

The 3×3 unitary matrix V, which actually links the neutrino mass eigenstates (ν_1,ν_2,ν_3) to the neutrino flavor eigenstates (ν_e,ν_μ,ν_τ), just measures the phenomenon of neutrino mixing.

A salient feature of massive Dirac neutrinos is lepton number conservation. To see why massive Dirac neutrinos are lepton-number-conserving, we make the global phase transformations

$$l(x) \to e^{i\Phi}l(x) \; , \quad \nu'_{\rm L}(x) \to e^{i\Phi}\nu'_{\rm L}(x) \; , \quad N'_{\rm R}(x) \to e^{i\Phi}N'_{\rm R}(x) \; , \tag{13}$$

where l denotes the column vector of e, μ and τ fields, and Φ is an arbitrary spacetime-independent phase. As the mass term $\mathcal{L}'_{\rm Dirac}$, the kinetic term $\mathcal{L}_{\rm kinetic}$ and the charged-current interaction term $\mathcal{L}_{\rm cc}$ are all invariant under these transformations, the lepton number must be conserved for massive Dirac neutrinos. It is evident that lepton flavors are violated, unless $M_{\rm D}$ is diagonal or equivalently V is the identity matrix. In other words, lepton flavor mixing leads to lepton flavor violation, or vice versa.

For example, the decay mode $\pi^- \to \mu^- + \overline{\nu}_\mu$ preserves both the lepton number and lepton flavors. In contrast, $\mu^+ \to e^+ + \gamma$ preserves the lepton number but violates the lepton flavors. The observed phenomena of neutrino oscillations verify the existence of neutrino flavor violation. Note that the $0\nu2\beta$ decay $(A,Z) \to (A,Z+2)+2e^-$ violates the lepton number. This process cannot take place if neutrinos are massive Dirac particles, but it may naturally happen if neutrinos are massive Majorana particles.

1.3 Majorana neutrino masses

The left-handed neutrino field $\nu_{\rm L}$ and its charge-conjugate counterpart $(\nu_{\rm L})^c$ can in principle form a neutrino mass term, as $(\nu_{\rm L})^c$ is actually right-handed. But this Majorana mass term is forbidden by the $SU(2)_{\rm L}\times U(1)_{\rm Y}$ gauge symmetry in the SM, which contains only one $SU(2)_{\rm L}$ Higgs doublet and preserves lepton number conservation. We shall show later that the introduction of an $SU(2)_{\rm L}$ Higgs triplet into the SM can accommodate such a neutrino mass term with gauge invariance. Here we ignore the details of the Higgs triplet models and focus on the Majorana neutrino mass term itself:

$$-\mathcal{L}'_{\rm Majorana} = \frac{1}{2}\overline{\nu_{\rm L}}M_{\rm L}(\nu_{\rm L})^c + {\rm h.c.} \; . \tag{14}$$

Note that the mass matrix $M_{\rm L}$ must be symmetric. Because the mass term is a Lorentz scalar whose transpose keeps unchanged, we have

$$\overline{\nu_L}M_{\rm L}(\nu_L)^c = \left[\overline{\nu_L}M_{\rm L}(\nu_L)^c\right]^T = -\overline{\nu_L}\mathcal{C}^T M_{\rm L}^T\overline{\nu_L}^T = \overline{\nu_L}M_{\rm L}^T(\nu_L)^c \; , \tag{15}$$

where a minus sign appears when interchanging two fermion field operators, and $\mathcal{C}^T = -\mathcal{C}$ has been used. Hence $M_{\rm L}^T = M_{\rm L}$ holds. This symmetric mass matrix can be diagonalized by the transformation $V^\dagger M_{\rm L} V^* = \widehat{M}_\nu \equiv {\rm Diag}\{m_1,m_2,m_3\}$, where V is a unitary matrix. After this, Eq. (14) becomes

$$-\mathcal{L}'_{\rm Majorana} = \frac{1}{2}\overline{\nu'_{\rm L}}\widehat{M}_\nu(\nu'_{\rm L})^c + {\rm h.c.} \; , \tag{16}$$

where $\nu'_{\rm L} = V^\dagger \nu_{\rm L}$ and $(\nu'_{\rm L})^c = \mathcal{C}\overline{\nu'_{\rm L}}^T$. Then the Majorana field

$$\nu' = \nu'_{\rm L} + (\nu'_{\rm L})^c = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} \; , \tag{17}$$

which certainly satisfies the Majorana condition $(\nu')^c = \nu'$, describes the mass eigenstates of three Majorana neutrinos. In other words,

$$-\mathcal{L}'_{\rm Majorana} = \frac{1}{2}\overline{\nu'}\widehat{M}_\nu \nu' = \frac{1}{2}\sum_{i=1}^{3} m_i \overline{\nu_i}\nu_i \; . \tag{18}$$

The kinetic term of Majorana neutrinos reads

$$\mathcal{L}_{\rm kinetic} = i\overline{\nu_{\rm L}}\gamma_\mu \partial^\mu \nu_{\rm L} = i\overline{\nu'_{\rm L}}\gamma_\mu \partial^\mu \nu'_{\rm L} = \frac{i}{2}\overline{\nu'}\gamma_\mu \partial^\mu \nu' = \frac{i}{2}\sum_{k=1}^{3} \overline{\nu_k}\gamma_\mu \partial^\mu \nu_k \; , \tag{19}$$

where we have used a generic relationship $\overline{(\psi_{\rm L})^c}\gamma_\mu \partial^\mu (\psi_{\rm L})^c = \overline{\psi_{\rm L}}\gamma_\mu \partial^\mu \psi_{\rm L}$. This relationship can easily be proved by taking account of $\partial^\mu \left[\overline{(\psi_{\rm L})^c}\gamma_\mu (\psi_{\rm L})^c\right] = 0$; i.e., we have

$$\begin{aligned} \overline{(\psi_{\rm L})^c}\gamma_\mu \partial^\mu (\psi_{\rm L})^c &= -\partial^\mu \overline{(\psi_{\rm L})^c}\gamma_\mu (\psi_{\rm L})^c = -\left[\partial^\mu \overline{(\psi_{\rm L})^c}\gamma_\mu (\psi_{\rm L})^c\right]^T \\ &= \left(\mathcal{C}\overline{\psi_{\rm L}}^T\right)^T \gamma_\mu^T \partial^\mu \left[(\psi_{\rm L})^T \mathcal{C}\right]^T = \overline{\psi_{\rm L}}\gamma_\mu \partial^\mu \psi_{\rm L}, \end{aligned} \tag{20}$$

where $\mathcal{C}^T \gamma_\mu^T \mathcal{C}^T = \gamma_\mu$, which may be read off from Eq. (4), has been used.

It is worth pointing out that the factor $1/2$ in $\mathcal{L}'_{\rm Majorana}$ allows us to get the Dirac equation of massive Majorana neutrinos analogous to that of massive Dirac neutrinos. To see this point more clearly, let us consider the Lagrangian of free Majorana neutrinos (i.e., their kinetic and mass terms):

$$\begin{aligned} \mathcal{L}_\nu &= i\overline{\nu_{\rm L}}\gamma_\mu \partial^\mu \nu_{\rm L} - \left[\frac{1}{2}\overline{\nu_{\rm L}} M_{\rm L} (\nu_{\rm L})^c + {\rm h.c.}\right] = i\overline{\nu'_{\rm L}}\gamma_\mu \partial^\mu \nu'_{\rm L} - \left[\frac{1}{2}\overline{\nu'_{\rm L}}\widehat{M}_\nu (\nu'_{\rm L})^c + {\rm h.c.}\right] \\ &= \frac{1}{2}\left(i\overline{\nu'}\gamma_\mu \partial^\mu \nu' - \overline{\nu'}\widehat{M}_\nu \nu'\right) = -\frac{1}{2}\left(i\partial^\mu \overline{\nu'}\gamma_\mu \nu' + \overline{\nu'}\widehat{M}_\nu \nu'\right) \; , \end{aligned} \tag{21}$$

where $\partial^\mu(\overline{\nu'}\gamma_\mu \nu') = 0$ has been used. Then we substitute $\mathcal{L}_\nu$ into the Euler-Lagrange equation

$$\partial^\mu \frac{\partial \mathcal{L}_\nu}{\partial\left(\partial^\mu \overline{\nu'}\right)} - \frac{\partial \mathcal{L}_\nu}{\partial \overline{\nu'}} = 0 \tag{22}$$

and obtain the Dirac equation

$$i\gamma_\mu \partial^\mu \nu' - \widehat{M}_\nu \nu' = 0 \; . \tag{23}$$

More explicitly, $i\gamma_\mu \partial^\mu \nu_k - m_k \nu_k = 0$ holds (for $k = 1, 2, 3$). That is why the factor $1/2$ in $\mathcal{L}'_{\rm Majorana}$ makes sense.

The weak charged-current interactions of three neutrinos in Eq. (5) can now be rewritten in terms of their mass eigenstates $\nu'_{\rm L} = V^\dagger \nu_{\rm L}$. In the chosen basis where the flavor and mass eigenstates of three charged leptons are identical, the expression of $\mathcal{L}_{cc}$ for Majorana neutrinos is the same as that given in Eq. (12) for Dirac neutrinos. The unitary matrix V is just the 3×3 Majorana neutrino mixing matrix, which contains two more irremovable CP-violating phases than the 3×3 Dirac neutrino mixing matrix (see section 4 for detailed discussions).

The most salient feature of massive Majorana neutrinos is lepton number violation. Let us make the global phase transformations

$$l(x) \to e^{i\Phi} l(x) \; , \quad \nu'_{\rm L}(x) \to e^{i\Phi} \nu'_{\rm L}(x) \; , \tag{24}$$

where l stands for the column vector of e, μ and τ fields, and Φ is an arbitrary spacetime-independent phase. One can immediately see that the kinetic term $\mathcal{L}_{\rm kinetic}$ and the charged-current interaction term $\mathcal{L}_{\rm cc}$ are invariant under these transformations, but the mass term $\mathcal{L}'_{\rm Majorana}$ is not invariant because of both $\overline{\nu'_{\rm L}} \to e^{-i\Phi} \overline{\nu'_{\rm L}}$ and $(\nu'_{\rm L})^c \to e^{-i\Phi} (\nu'_{\rm L})^c$. The lepton number is therefore violated for massive Majorana neutrinos. Similar to the case of Dirac neutrinos, the lepton flavor violation of Majorana neutrinos is also described by V.

The $0\nu2\beta$ decay $(A, Z) \to (A, Z+2) + 2e^-$ is a clean signature of the Majorana nature of massive neutrinos. This lepton-number-violating process can occur when there exists neutrino-antineutrino mixing induced by the Majorana mass term (i.e., the neutrino mass eigenstates are self-conjugate, $\overline{\nu}_i = \nu_i$). The effective mass of the $0\nu2\beta$ decay is defined as

$$\langle m \rangle_{ee} \equiv \left| \sum_i m_i V_{ei}^2 \right| \; , \tag{25}$$

where m_i comes from the helicity suppression factor m_i/E for the ν_i exchange between two beta decays with E being the energy of the virtual ν_i neutrino. Current experimental data only yield an upper bound $\langle m \rangle_{ee} < 0.23$ eV (or < 0.85 eV as a more conservative bound) at the 2σ level.

1.4 Hybrid neutrino mass terms

Similar to Eq. (14), the right-handed neutrino field $N_{\rm R}$ and its charge-conjugate counterpart $(N_{\rm R})^c$ can also form a Majorana mass term. Hence it is possible to write out the following hybrid neutrino mass terms in terms of $\nu_{\rm L}$, $N_{\rm R}$, $(\nu_{\rm L})^c$ and $(N_{\rm R})^c$ fields:

$$\begin{aligned} -\mathcal{L}'_{\rm hybrid} &= \overline{\nu_{\rm L}} M_{\rm D} N_{\rm R} + \frac{1}{2} \overline{\nu_{\rm L}} M_{\rm L} (\nu_{\rm L})^c + \frac{1}{2} \overline{(N_{\rm R})^c} M_{\rm R} N_{\rm R} + {\rm h.c.} \\ &= \frac{1}{2} \left[\overline{\nu_{\rm L}} \quad \overline{(N_{\rm R})^c} \right] \begin{pmatrix} M_{\rm L} & M_{\rm D} \\ M_{\rm D}^T & M_{\rm R} \end{pmatrix} \begin{bmatrix} (\nu_{\rm L})^c \\ N_{\rm R} \end{bmatrix} + {\rm h.c.} \; , \end{aligned} \tag{26}$$

where $M_{\rm L}$ and $M_{\rm R}$ are symmetric mass matrices because the corresponding mass terms are of the Majorana type, and the relationship

$$\overline{(N_{\rm R})^c} M_{\rm D}^T (\nu_{\rm L})^c = \left[(N_{\rm R})^T \mathcal{C} M_{\rm D}^T \mathcal{C} \overline{\nu_{\rm L}}^T \right]^T = \overline{\nu_{\rm L}} M_{\rm D} N_{\rm R} \tag{27}$$

has been used. The overall 6×6 mass matrix in Eq. (26) is also symmetric, and thus it can be diagonalized by a 6×6 unitary matrix through the transformation

$$\begin{pmatrix} V & R \\ S & U \end{pmatrix}^\dagger \begin{pmatrix} M_{\rm L} & M_{\rm D} \\ M_{\rm D}^T & M_{\rm R} \end{pmatrix} \begin{pmatrix} V & R \\ S & U \end{pmatrix}^* = \begin{pmatrix} \widehat{M}_\nu & \mathbf{0} \\ \mathbf{0} & \widehat{M}_N \end{pmatrix} \; , \tag{28}$$

where we have defined $\widehat{M}_\nu \equiv {\rm Diag}\{m_1, m_2, m_3\}$, $\widehat{M}_N \equiv {\rm Diag}\{M_1, M_2, M_3\}$, and the 3×3 matrices V, R, S and U satisfy the unitarity conditions

$$\begin{aligned} V V^\dagger + R R^\dagger = S S^\dagger + U U^\dagger &= \mathbf{1} \; , \\ V^\dagger V + S^\dagger S = R^\dagger R + U^\dagger U &= \mathbf{1} \; , \\ V S^\dagger + R U^\dagger = V^\dagger R + S^\dagger U &= \mathbf{0} \; . \end{aligned} \tag{29}$$

After this diagonalization, Eq. (26) becomes

$$-\mathcal{L}'_{\rm hybrid} = \frac{1}{2} \left[\overline{\nu'_{\rm L}} \quad \overline{(N'_{\rm R})^c} \right] \begin{pmatrix} \widehat{M}_\nu & \mathbf{0} \\ \mathbf{0} & \widehat{M}_N \end{pmatrix} \begin{bmatrix} (\nu'_{\rm L})^c \\ N'_{\rm R} \end{bmatrix} + {\rm h.c.} \; , \tag{30}$$

where $\nu'_{\rm L} = V^\dagger \nu_{\rm L} + S^\dagger (N_{\rm R})^c$ and $N'_{\rm R} = R^T (\nu_{\rm L})^c + U^T N_{\rm R}$ together with $(\nu'_{\rm L})^c = \mathcal{C} \overline{\nu'_{\rm L}}^T$ and $(N'_{\rm R})^c = \mathcal{C} \overline{N'_{\rm R}}^T$. Then the Majorana field

$$\nu' = \begin{bmatrix} \nu'_{\rm L} \\ (N'_{\rm R})^c \end{bmatrix} + \begin{bmatrix} (\nu'_{\rm L})^c \\ N'_{\rm R} \end{bmatrix} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ N_1 \\ N_2 \\ N_3 \end{pmatrix} \tag{31}$$

satisfies the Majorana condition $(\nu')^c = \nu'$ and describes the mass eigenstates of six Majorana neutrinos. In other words,

$$-\mathcal{L}'_{\rm hybrid} = \frac{1}{2} \overline{\nu'} \begin{pmatrix} \widehat{M}_\nu & \mathbf{0} \\ \mathbf{0} & \widehat{M}_N \end{pmatrix} \nu' = \frac{1}{2} \sum_{i=1}^{3} \left(m_i \overline{\nu_i} \nu_i + M_i \overline{N_i} N_i \right) \; . \tag{32}$$

Because of $\nu_{\rm L} = V \nu'_{\rm L} + R (N'_{\rm R})^c$ and $N_{\rm R} = S^* (\nu'_{\rm L})^c + U^* N'_{\rm R}$, we immediately have $(\nu_{\rm L})^c = V^* (\nu'_{\rm L})^c + R^* N'_{\rm R}$ and $(N_{\rm R})^c = S \nu'_{\rm L} + U (N'_{\rm R})^c$. Given the generic relations $\overline{(\psi_{\rm L})^c} \gamma_\mu \partial^\mu (\psi_{\rm L})^c = \overline{\psi_{\rm L}} \gamma_\mu \partial^\mu \psi_{\rm L}$ and $\overline{(\psi_{\rm R})^c} \gamma_\mu \partial^\mu (\psi_{\rm R})^c = \overline{\psi_{\rm R}} \gamma_\mu \partial^\mu \psi_{\rm R}$ for an arbitrary fermion field ψ, the kinetic term of Majorana neutrinos under consideration turns out to be

$$\begin{aligned} \mathcal{L}_{\rm kinetic} &= i \overline{\nu_{\rm L}} \gamma_\mu \partial^\mu \nu_{\rm L} + i \overline{N_{\rm R}} \gamma_\mu \partial^\mu N_{\rm R} = i \overline{\nu'_{\rm L}} \gamma_\mu \partial^\mu \nu'_{\rm L} + i \overline{N'_{\rm R}} \gamma_\mu \partial^\mu N'_{\rm R} = \frac{i}{2} \overline{\nu'} \gamma_\mu \partial^\mu \nu' \\ &= \frac{i}{2} \sum_{k=1}^{3} \left(\overline{\nu_k} \gamma_\mu \partial^\mu \nu_k + \overline{N_k} \gamma_\mu \partial^\mu N_k \right) \; , \end{aligned} \tag{33}$$

where the unitarity conditions given in Eq. (29) have been used.

The weak charged-current interactions of active neutrinos in Eq. (5) can now be rewritten in terms of the mass eigenstates of six Majorana neutrinos via $\nu_{\rm L} = V \nu'_{\rm L} + R (N'_{\rm R})^c$. In the chosen basis where the flavor and mass eigenstates of three charged leptons are identical, we have

$$\mathcal{L}_{\rm cc} = \frac{g}{\sqrt{2}} \overline{(e \; \mu \; \tau)_{\rm L}} \; \gamma^\mu \left[V \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix}_{\rm L} + R \begin{pmatrix} N_1 \\ N_2 \\ N_3 \end{pmatrix}_{\rm L} \right] W^-_\mu + {\rm h.c.} \; . \tag{34}$$

Note that V and R are responsible for the charged-current interactions of three known neutrinos ν_i and three new neutrinos N_i (for $i = 1, 2, 3$), respectively. Their correlation is described by $V V^\dagger + R R^\dagger = \mathbf{1}$, and thus V is not unitary unless ν_i and N_i are completely decoupled (i.e., $R = \mathbf{0}$).

As a consequence of lepton number violation, the $0\nu2\beta$ decay $(A, Z) \to (A, Z + 2) + 2e^-$ can now take place via the exchanges of both ν_i and N_i between two beta decays, whose coupling matrix elements are V_{ei} and R_{ei} respectively. The relative contributions of ν_i and N_i to this lepton-number-violating process depend not only on m_i, M_i, V_{ei} and R_{ei} but also on the relevant nuclear matrix elements which cannot be reliably evaluated. For a realistic seesaw mechanism working at the TeV scale (i.e., $M_i \sim \mathcal{O}(1)$ TeV) or at a superhigh-energy scale, however, the contribution of ν_i to the $0\nu2\beta$ decay is in most cases dominant.

The hybrid neutrino mass terms in Eq. (26) provide us with the necessary ingredients of a dynamic mechanism to interpret why three known neutrinos have non-zero but tiny masses. The key point is that the mass scales of $M_{\rm L}$, $M_{\rm D}$ and $M_{\rm R}$ may have a strong hierarchy. First, $M_{\rm D} \sim \langle H \rangle \approx 174$ GeV is naturally characterized by the electroweak symmetry breaking scale. Second, $M_{\rm L} \ll \langle H \rangle$ satisfies 't Hooft's naturalness criterion because this Majorana mass term violates lepton number conservation. Third, $M_{\rm R} \gg \langle H \rangle$ is naturally expected since right-handed neutrinos are $SU(2)_{\rm L}$ gauge singlets and thus their mass term is not subject to the electroweak symmetry breaking scale. The hierarchy $M_{\rm R} \gg M_{\rm D} \gg M_{\rm L}$ can therefore allow us to make reliable approximations in deriving the effective mass matrix of three active neutrinos $(\nu_e, \nu_\mu, \nu_\tau)$ from Eq. (28). The latter yields

$$\begin{aligned} R\widehat{M}_N &= M_{\rm L} R^* + M_{\rm D} U^* \; , \\ S\widehat{M}_\nu &= M_{\rm D}^T V^* + M_{\rm R} S^* \; ; \end{aligned} \tag{35}$$

and

$$\begin{aligned} U\widehat{M}_N &= M_{\rm R} U^* + M_{\rm D}^T R^* \; , \\ V\widehat{M}_\nu &= M_{\rm L} V^* + M_{\rm D} S^* \; . \end{aligned} \tag{36}$$

Given $M_{\rm R} \gg M_{\rm D} \gg M_{\rm L}$, $R \sim S \sim \mathcal{O}(M_{\rm D}/M_{\rm R})$ naturally holds, implying that U and V are almost unitary up to the accuracy of $\mathcal{O}(M_{\rm D}^2/M_{\rm R}^2)$. Hence Eq. (36) leads to

$$\begin{aligned} U\widehat{M}_N U^T &= M_{\rm R} (U U^\dagger)^T + M_{\rm D}^T (R^* U^T) \approx M_{\rm R} \; , \\ V\widehat{M}_\nu V^T &= M_{\rm L} (V V^\dagger)^T + M_{\rm D} (S^* V^T) \approx M_{\rm L} + M_{\rm D} (S^* V^T) \; . \end{aligned} \tag{37}$$

$S^* V^T = M_{\rm R}^{-1} S \widehat{M}_\nu V^T - M_{\rm R}^{-1} M_{\rm D}^T (V V^\dagger)^T \approx -M_{\rm R}^{-1} M_{\rm D}^T$ can be derived from Eq. (35). We substitute this expression into Eq. (37) and then obtain

$$M_\nu \equiv V \widehat{M}_\nu V^T \approx M_{\rm L} - M_{\rm D} M_{\rm R}^{-1} M_{\rm D}^T \; . \tag{38}$$

This result, known as the type-(I+II) seesaw relation, is just the effective mass matrix of three light neutrinos. The small mass scale of M_ν is attributed to the small mass scale of $M_{\rm L}$ and the large mass scale of $M_{\rm R}$. There are two particularly interesting limits: (1) If $M_{\rm L}$ is absent from Eq. (26), one will be left with the canonical or type-I seesaw relation $M_\nu \approx -M_{\rm D} M_{\rm R}^{-1} M_{\rm D}^T$; (2) If only $M_{\rm L}$ is present in Eq. (26), one will get the type-II seesaw relation $M_\nu = M_{\rm L}$. More detailed discussions about various seesaw mechanisms and their phenomenological consequences will be presented in sections 6, 7 and 8.

2 Diagnosis of CP Violation

2.1 C, P and T transformations

We begin with a brief summary of the transformation properties of quantum fields under the discrete space-time symmetries of parity (P), charge conjugation (C) and time reversal (T). The parity transformation changes the space coordinates $\vec{x}$ into $-\vec{x}$. The charge conjugation flips the signs of internal charges of a particle, such as the electric charge and the lepton (baryon) number. The time reversal reflects the time coordinate t into $-t$.

A free Dirac spinor $\psi(t, \vec{x})$ or $\overline{\psi}(t, \vec{x})$ transforms under C, P and T as [3]

$$\begin{aligned} \psi(t, \vec{x}) &\xrightarrow{\rm C} \mathcal{C} \overline{\psi}^T(t, \vec{x}) \; , \\ \overline{\psi}(t, \vec{x}) &\xrightarrow{\rm C} -\psi^T(t, \vec{x}) \mathcal{C}^{-1} \; , \end{aligned}$$

[3] For simplicity, here we have omitted a phase factor associated with each transformation. Because one is always interested in the spinor bilinears, the relevant phase factor usually plays no physical role.

Table 1: Transformation properties of the scalar-, pseudoscalar-, vector-, pseudovector- and tensor-like spinor bilinears under C, P and T. Here $\vec{x} \to -\vec{x}$ under P, CP and CPT, together with $t \to -t$ under T and CPT, is hidden and self-explaining for ψ_1 and ψ_2.

	$\overline{\psi_1}\psi_2$	$i\overline{\psi_1}\gamma_5\psi_2$	$\overline{\psi_1}\gamma_\mu\psi_2$	$\overline{\psi_1}\gamma_\mu\gamma_5\psi_2$	$\overline{\psi_1}\sigma_{\mu\nu}\psi_2$
C	$\overline{\psi_2}\psi_1$	$i\overline{\psi_2}\gamma_5\psi_1$	$-\overline{\psi_2}\gamma_\mu\psi_1$	$\overline{\psi_2}\gamma_\mu\gamma_5\psi_1$	$-\overline{\psi_2}\sigma_{\mu\nu}\psi_1$
P	$\overline{\psi_1}\psi_2$	$-i\overline{\psi_1}\gamma_5\psi_2$	$\overline{\psi_1}\gamma^\mu\psi_2$	$-\overline{\psi_1}\gamma^\mu\gamma_5\psi_2$	$\overline{\psi_1}\sigma^{\mu\nu}\psi_2$
T	$\overline{\psi_1}\psi_2$	$-i\overline{\psi_1}\gamma_5\psi_2$	$\overline{\psi_1}\gamma^\mu\psi_2$	$\overline{\psi_1}\gamma^\mu\gamma_5\psi_2$	$-\overline{\psi_1}\sigma^{\mu\nu}\psi_2$
CP	$\overline{\psi_2}\psi_1$	$-i\overline{\psi_2}\gamma_5\psi_1$	$-\overline{\psi_2}\gamma^\mu\psi_1$	$-\overline{\psi_2}\gamma^\mu\gamma_5\psi_1$	$-\overline{\psi_2}\sigma^{\mu\nu}\psi_1$
CPT	$\overline{\psi_2}\psi_1$	$i\overline{\psi_2}\gamma_5\psi_1$	$-\overline{\psi_2}\gamma_\mu\psi_1$	$-\overline{\psi_2}\gamma_\mu\gamma_5\psi_1$	$\overline{\psi_2}\sigma_{\mu\nu}\psi_1$

$$\begin{aligned}
\psi(t,\vec{x}) &\xrightarrow{\rm P} \mathcal{P}\psi(t,-\vec{x}) \; , \\
\overline{\psi}(t,\vec{x}) &\xrightarrow{\rm P} \overline{\psi}(t,-\vec{x})\mathcal{P}^\dagger \; , \\
\psi(t,\vec{x}) &\xrightarrow{\rm T} \mathcal{T}\psi(-t,\vec{x}) \; , \\
\overline{\psi}(t,\vec{x}) &\xrightarrow{\rm T} \overline{\psi}(-t,\vec{x})\mathcal{T}^\dagger \; ,
\end{aligned} \tag{39}$$

where $\mathcal{C} = i\gamma_2\gamma_0$, $\mathcal{P} = \gamma_0$ and $\mathcal{T} = \gamma_1\gamma_3$ in the Dirac-Pauli representation. These transformation properties can simply be deduced from the requirement that the Dirac equation $i\gamma_\mu\partial^\mu\psi(t,\vec{x}) = m\psi(t,\vec{x})$ be invariant under C, P or T operation. Note that all the classical numbers (or c-numbers), such as the coupling constants and γ-matrix elements, must be complex-conjugated under T. Note also that the charge-conjugation matrix $\mathcal{C}$ satisfies the conditions given in Eq. (4). It is very important to figure out how the Dirac spinor bilinears transform under C, P and T, because both leptons and quarks are described by spinor fields and they always appear in the bilinear forms in a Lorentz-invariant Lagrangian. Let us consider the following scalar-, pseudoscalar-, vector-, pseudovector- and tensor-like spinor bilinears: $\overline{\psi_1}\psi_2$, $i\overline{\psi_1}\gamma_5\psi_2$, $\overline{\psi_1}\gamma_\mu\psi_2$, $\overline{\psi_1}\gamma_\mu\gamma_5\psi_2$ and $\overline{\psi_1}\sigma_{\mu\nu}\psi_2$, where $\sigma_{\mu\nu} \equiv i[\gamma_\mu,\gamma_\nu]/2$ is defined. One may easily verify that all these bilinears are Hermitian. Under C, P and T, for example,

$$\begin{aligned}
\overline{\psi_1}\gamma_\mu\psi_2 \xrightarrow{\rm C} -\psi_1^T\mathcal{C}^{-1}\gamma_\mu\mathcal{C}\overline{\psi_2}^T = \psi_1^T\gamma_\mu^T\overline{\psi_2}^T = -\left[\overline{\psi_2}\gamma_\mu\psi_1\right]^T = -\overline{\psi_2}\gamma_\mu\psi_1 \; , \\
\overline{\psi_1}\gamma_\mu\psi_2 \xrightarrow{\rm P} \overline{\psi_1}\gamma_0\gamma_\mu\gamma_0\psi_2 = \overline{\psi_1}\gamma^\mu\psi_2 \; , \\
\overline{\psi_1}\gamma_\mu\psi_2 \xrightarrow{\rm T} \overline{\psi_1}\left(\gamma_1\gamma_3\right)^\dagger\gamma_\mu^*\left(\gamma_1\gamma_3\right)\psi_2 = \overline{\psi_1}\gamma^\mu\psi_2 \; ;
\end{aligned} \tag{40}$$

and thus

$$\begin{aligned}
\overline{\psi_1}\gamma_\mu\psi_2 &\xrightarrow{\rm CP} -\overline{\psi_2}\gamma^\mu\psi_1 \; , \\
\overline{\psi_1}\gamma_\mu\psi_2 &\xrightarrow{\rm CPT} -\overline{\psi_2}\gamma_\mu\psi_1 \; ,
\end{aligned} \tag{41}$$

with $\vec{x} \to -\vec{x}$ under P and $t \to -t$ under T for ψ_1 and ψ_2. The transformation properties of five spinor bilinears under C, P, T, CP and CPT are summarized in Table 1, where one should keep in mind that all the c-numbers are complex-conjugated under T and CPT.

It is well known that CPT is a good symmetry in a local quantum field theory which is Lorentz-invariant and possesses a Hermitian Lagrangian. The latter is necessary in order to have a unitary transition operator (i.e., the S-matrix). The CPT invariance of a theory implies that CP and T must be simultaneously conserving or broken, as already examined in the quark sector of the SM via the K^0-$\bar{K}^0$ mixing system. After a slight modification of the SM by introducing the Dirac or Majorana mass term for three neutrinos, one may also look at possible sources of CP or T violation in the lepton sector.

2.2 The source of CP violation

The SM of electroweak interactions is based on the $SU(2)_{\rm L} \times U(1)_{\rm Y}$ gauge symmetry and the Higgs mechanism. The latter triggers the spontaneous symmetry breaking $SU(2)_{\rm L} \times U(1)_{\rm Y} \to U(1)_{\rm Q}$, such that three gauge bosons, three charged leptons and six quarks can all acquire masses. But this mechanism itself does not spontaneously break CP, and thus one may examine the source of CP violation in the SM either before or after spontaneous symmetry breaking.

The Lagrangian of the SM $\mathcal{L} = \mathcal{L}_{\rm G} + \mathcal{L}_{\rm H} + \mathcal{L}_{\rm F} + \mathcal{L}_{\rm Y}$ is composed of four parts: the kinetic term of the gauge fields and their self-interactions ($\mathcal{L}_{\rm G}$), the kinetic term of the Higgs doublet and its potential and interactions with the gauge fields ($\mathcal{L}_{\rm H}$), the kinetic term of the fermion fields and their interactions with the gauge fields ($\mathcal{L}_{\rm F}$), and the Yukawa interactions of the fermion fields with the Higgs doublet ($\mathcal{L}_{\rm Y}$):

$$\begin{aligned}
\mathcal{L}_{\rm G} &= -\frac{1}{4}\left(W^{i\mu\nu}W^i_{\mu\nu} + B^{\mu\nu}B_{\mu\nu}\right) \ , \\
\mathcal{L}_{\rm H} &= \left(D^\mu H\right)^\dagger \left(D_\mu H\right) - \mu^2 H^\dagger H - \lambda\left(H^\dagger H\right)^2 \ , \\
\mathcal{L}_{\rm F} &= \overline{Q_{\rm L}} i\not{D} Q_{\rm L} + \overline{\ell_{\rm L}} i\not{D} \ell_{\rm L} + \overline{U_{\rm R}} i\not{\partial}' U_{\rm R} + \overline{D_{\rm R}} i\not{\partial}' D_{\rm R} + \overline{E_{\rm R}} i\not{\partial}' E_{\rm R} \ , \\
\mathcal{L}_{\rm Y} &= -\overline{Q_{\rm L}} Y_{\rm u} \tilde{H} U_{\rm R} - \overline{Q_{\rm L}} Y_{\rm d} H D_{\rm R} - \overline{\ell_{\rm L}} Y_l H E_{\rm R} + {\rm h.c.} \ ,
\end{aligned} \tag{42}$$

whose notations are self-explanatory. To accommodate massive neutrinos, the simplest way is to slightly modify the $\mathcal{L}_{\rm F}$ and $\mathcal{L}_{\rm Y}$ parts (e.g., by introducing three right-handed neutrinos into the SM and allowing for the Yukawa interactions between neutrinos and the Higgs doublet). CP violation is due to the coexistence of $\mathcal{L}_{\rm F}$ and $\mathcal{L}_{\rm Y}$.

We first show that $\mathcal{L}_{\rm G}$ is always invariant under CP. The transformation properties of gauge fields B_μ and W^i_μ under C and P are

$$\begin{aligned}
\left[B_\mu,\, W^1_\mu,\, W^2_\mu,\, W^3_\mu\right] &\xrightarrow{\rm C} \left[-B_\mu,\, -W^1_\mu,\, +W^2_\mu,\, -W^3_\mu\right] \ , \\
\left[B_\mu,\, W^1_\mu,\, W^2_\mu,\, W^3_\mu\right] &\xrightarrow{\rm P} \left[B^\mu,\, W^{1\mu},\, W^{2\mu},\, W^{3\mu}\right] \ , \\
\left[B_\mu,\, W^1_\mu,\, W^2_\mu,\, W^3_\mu\right] &\xrightarrow{\rm CP} \left[-B^\mu,\, -W^{1\mu},\, +W^{2\mu},\, -W^{3\mu}\right]
\end{aligned} \tag{43}$$

with $\vec{x} \to -\vec{x}$ under P and CP for relevant fields. Then the gauge field tensors $B_{\mu\nu}$ and $W^i_{\mu\nu}$ transform under CP as follows:

$$\left[B_{\mu\nu},\, W^1_{\mu\nu},\, W^2_{\mu\nu},\, W^3_{\mu\nu}\right] \xrightarrow{\rm CP} \left[-B^{\mu\nu},\, -W^{1\mu\nu},\, +W^{2\mu\nu},\, -W^{3\mu\nu}\right] \ . \tag{44}$$

Hence $\mathcal{L}_{\rm G}$ is formally invariant under CP.

We proceed to show that $\mathcal{L}_{\rm H}$ is also invariant under CP. The Higgs doublet H contains two scalar components ϕ^+ and ϕ^0; i.e.,

$$H = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \ , \quad H^\dagger = \begin{pmatrix} \phi^- & \phi^{0*} \end{pmatrix} \ . \tag{45}$$

Therefore,

$$H(t, \vec{x}) \xrightarrow{\rm CP} H^*(t, -\vec{x}) = \begin{pmatrix} \phi^- \\ \phi^{0*} \end{pmatrix} \ . \tag{46}$$

It is very trivial to prove that the $H^\dagger H$ and $(H^\dagger H)^2$ terms of $\mathcal{L}_{\rm H}$ are CP-invariant. To examine how the $(D^\mu H)^\dagger (D_\mu H)$ term of $\mathcal{L}_{\rm H}$ transforms under CP, we explicitly write out

$$D_\mu H = \left(\partial_\mu - ig\tau^k W^k_\mu - ig' Y B_\mu\right) H = \begin{pmatrix} \partial_\mu \phi^+ - iX^+_\mu \phi^0 - iY^+_\mu \phi^+ \\ \partial_\mu \phi^0 - iX^-_\mu \phi^+ + iY^-_\mu \phi^0 \end{pmatrix} \tag{47}$$

with $X^\pm_\mu \equiv gW^\pm_\mu/\sqrt{2} = g(W^1_\mu \mp iW^2_\mu)/2$, $Y^\pm \equiv \pm g' Y B_\mu + gW^3_\mu/2$, and $k = 1, 2, 3$. Note that

$$X^\pm_\mu \xrightarrow{\rm CP} -X^{\mp\mu}\; , \quad Y^\pm_\mu \xrightarrow{\rm CP} -Y^{\pm\mu}\; , \tag{48}$$

together with $\partial_\mu \to \partial^\mu$, $\phi^\pm \to \phi^\mp$ and $\phi^0 \to \phi^{0*}$ under CP. So it is easy to check that $(D^\mu H)^\dagger (D_\mu H)$ is also CP-invariant. Therefore, $\mathcal{L}_{\rm H}$ is formally invariant under CP.

The next step is to examine the CP invariance of $\mathcal{L}_{\rm F}$. To be more specific, we divide $\mathcal{L}_{\rm F}$ into the quark sector and the lepton sector; i.e., $\mathcal{L}_{\rm F} = \mathcal{L}_q + \mathcal{L}_l$. We only analyze the CP property of $\mathcal{L}_q$ in the following, because that of $\mathcal{L}_l$ can be analyzed in the same way. The explicit form of $\mathcal{L}_q$ reads

$$\begin{aligned}
\mathcal{L}_q = \overline{Q_{\rm L}} i\not{D} Q_{\rm L} + \overline{U_{\rm R}} i\not{\partial}' U_{\rm R} + \overline{D_{\rm R}} i\not{\partial}' D_{\rm R} = \sum_{j=1}^{3} \Big\{ & \frac{g}{2}\left[\overline{q'_j}\gamma^\mu P_{\rm L} W^1_\mu q_j + \overline{q_j}\gamma^\mu P_{\rm L} W^1_\mu q'_j\right] \\
& + \frac{g}{2}\left[i\overline{q'_j}\gamma^\mu P_{\rm L} W^2_\mu q_j - i\overline{q_j}\gamma^\mu P_{\rm L} W^2_\mu q'_j\right] \\
& + \frac{g}{2}\left[\overline{q_j}\gamma^\mu P_{\rm L} W^3_\mu q_j - \overline{q'_j}\gamma^\mu P_{\rm L} W^3_\mu q'_j\right] \\
& + i\left[\overline{q_j}\gamma^\mu P_{\rm L}\left(\partial_\mu - i\frac{g'}{6}B_\mu\right) q_j\right] \\
& + i\left[\overline{q'_j}\gamma^\mu P_{\rm L}\left(\partial_\mu - i\frac{g'}{6}B_\mu\right) q'_j\right] \\
& + i\left[\overline{q_j}\gamma^\mu P_{\rm R}\left(\partial_\mu - i\frac{2g'}{3}B_\mu\right) q_j\right] \\
& + i\left[\overline{q'_j}\gamma^\mu P_{\rm R}\left(\partial_\mu + i\frac{g'}{3}B_\mu\right) q'_j\right]\Big\}\; ,
\end{aligned} \tag{49}$$

where q_j and q'_j (for $j = 1, 2, 3$) run over (u, c, t) and (d, s, b), respectively. The transformation properties of gauge fields B_μ and W^i_μ under C and P have been given in Eq. (43). With the help of Table 1, one can see that the relevant spinor bilinears transform under C and P as follows:

$$\begin{aligned}
\overline{\psi_1}\gamma_\mu\left(1 \pm \gamma_5\right)\psi_2 &\xrightarrow{\rm C} -\overline{\psi_2}\gamma_\mu\left(1 \mp \gamma_5\right)\psi_1\; , \\
\overline{\psi_1}\gamma_\mu\left(1 \pm \gamma_5\right)\psi_2 &\xrightarrow{\rm P} +\overline{\psi_1}\gamma^\mu\left(1 \mp \gamma_5\right)\psi_2\; , \\
\overline{\psi_1}\gamma_\mu\left(1 \pm \gamma_5\right)\psi_2 &\xrightarrow{\rm CP} -\overline{\psi_2}\gamma^\mu\left(1 \pm \gamma_5\right)\psi_1\; ,
\end{aligned} \tag{50}$$

with $\vec{x} \to -\vec{x}$ under P and CP for ψ_1 and ψ_2. Furthermore,

$$\begin{aligned}
\overline{\psi_1}\gamma_\mu\left(1 \pm \gamma_5\right)\partial^\mu\psi_2 &\xrightarrow{\rm C} \overline{\psi_2}\gamma_\mu\left(1 \mp \gamma_5\right)\partial^\mu\psi_1\; , \\
\overline{\psi_1}\gamma_\mu\left(1 \pm \gamma_5\right)\partial^\mu\psi_2 &\xrightarrow{\rm P} \overline{\psi_1}\gamma^\mu\left(1 \mp \gamma_5\right)\partial_\mu\psi_2\; , \\
\overline{\psi_1}\gamma_\mu\left(1 \pm \gamma_5\right)\partial^\mu\psi_2 &\xrightarrow{\rm CP} \overline{\psi_2}\gamma^\mu\left(1 \pm \gamma_5\right)\partial_\mu\psi_1\; ,
\end{aligned} \tag{51}$$

with $\vec{x} \to -\vec{x}$ under P and CP for ψ_1 and ψ_2. It is straightforward to check that $\mathcal{L}_q$ in Eq. (49) is formally invariant under CP. Following the same procedure and using Eqs. (49), (50) and (51), one can easily show that $\mathcal{L}_l = \overline{\ell_{\rm L}} i\not{D} \ell_{\rm L} + \overline{E_{\rm R}} i\not{\partial}' E_{\rm R}$ is also CP-invariant. Thus we conclude that $\mathcal{L}_{\rm F}$ is invariant under CP.

The last step is to examine whether $\mathcal{L}_{\rm Y}$ is CP-conserving or not. Explicitly,

$$\begin{aligned}
-\mathcal{L}_{\rm Y} &= \overline{Q_{\rm L}} Y_{\rm u} \tilde{H} U_{\rm R} + \overline{Q_{\rm L}} Y_{\rm d} H D_{\rm R} + \overline{\ell_{\rm L}} Y_l H E_{\rm R} + {\rm h.c.} \\
&= \sum_{j,k=1}^{3} \Big\{ (Y_{\rm u})_{jk} \left[\overline{q_j} P_{\rm R} q_k \phi^{0*} - \overline{q'_j} P_{\rm R} q_k \phi^-\right]
\end{aligned}$$

$$
\begin{aligned}
&+(Y_{\rm u})^*_{jk}\left[\overline{q_k}P_{\rm L}q_j\phi^0-\overline{q_k}P_{\rm L}q'_j\phi^+\right]\\
&+(Y_{\rm d})_{jk}\left[\overline{q_j}P_{\rm R}q'_k\phi^+ +\overline{q'_j}P_{\rm R}q'_k\phi^0\right]\\
&+(Y_{\rm d})^*_{jk}\left[\overline{q'_k}P_{\rm L}q_j\phi^- +\overline{q'_k}P_{\rm L}q'_j\phi^{0^*}\right]\\
&+(Y_l)_{jk}\left[\overline{\nu_j}P_{\rm R}l_k\phi^+ +\overline{l_j}P_{\rm R}l_k\phi^0\right]\\
&+(Y_l)^*_{jk}\left[\overline{l_k}P_{\rm L}\nu_j\phi^- +\overline{l_k}P_{\rm L}l_j\phi^{0^*}\right]\Big\}\ , \qquad (52)
\end{aligned}
$$

where q_j and q'_j (for $j=1,2,3$) run over (u,c,t) and (d,s,b), respectively; while ν_j and l_j (for $j=1,2,3$) run over (ν_e,ν_μ,ν_τ) and (e,μ,τ), respectively. Because of $\phi^\pm\to\phi^\mp$, $\phi^0\to\phi^{0^*}$ and $\overline{\psi_1}(1\pm\gamma_5)\psi_2\to\overline{\psi_2}(1\mp\gamma_5)\psi_1$ under CP, we immediately arrive at

$$
\begin{aligned}
-\mathcal{L}_{\rm Y}\xrightarrow{\rm CP}\sum_{j,k=1}^{3}\Big\{&(Y_{\rm u})_{jk}\left[\overline{q_k}P_{\rm L}q_j\phi^0-\overline{q_k}P_{\rm L}q'_j\phi^+\right]\\
&+(Y_{\rm u})^*_{jk}\left[\overline{q_j}P_{\rm R}q_k\phi^{0^*}-\overline{q'_j}P_{\rm R}q_k\phi^-\right]\\
&+(Y_{\rm d})_{jk}\left[\overline{q'_k}P_{\rm L}q_j\phi^- +\overline{q'_k}P_{\rm L}q'_j\phi^{0^*}\right]\\
&+(Y_{\rm d})^*_{jk}\left[\overline{q_j}P_{\rm R}q'_k\phi^+ +\overline{q'_j}P_{\rm R}q'_k\phi^0\right]\\
&+(Y_l)_{jk}\left[\overline{l_k}P_{\rm L}\nu_j\phi^- +\overline{l_k}P_{\rm L}l_j\phi^{0^*}\right]\\
&+(Y_l)^*_{jk}\left[\overline{\nu_j}P_{\rm R}l_k\phi^+ +\overline{l_j}P_{\rm R}l_k\phi^0\right]\Big\}\ , \qquad (53)
\end{aligned}
$$

with $\vec{x}\to-\vec{x}$ for both scalar and spinor fields under consideration. Comparing between Eqs. (52) and (53), we see that $\mathcal{L}_{\rm Y}$ will be formally invariant under CP if the conditions

$$
(Y_{\rm u})_{jk}=(Y_{\rm u})^*_{jk}\ ,\quad (Y_{\rm d})_{jk}=(Y_{\rm d})^*_{jk}\ ,\quad (Y_l)_{jk}=(Y_l)^*_{jk} \qquad (54)
$$

are satisfied. In other words, the Yukawa coupling matrices $Y_{\rm u}$, $Y_{\rm d}$ and Y_l must be real to guarantee the CP invartiance of $\mathcal{L}_{\rm Y}$. Given three massless neutrinos in the SM, it is always possible to make Y_l real by redefining the phases of charged-lepton fields. But it is in general impossible to make both $Y_{\rm u}$ and $Y_{\rm d}$ real for three families of quarks, and thus CP violation can only appear in the quark sector.

Given massive neutrinos beyond the SM, $\mathcal{L}_{\rm Y}$ must be modified. The simplest way is to introduce three right-handed neutrinos and incorporate the Dirac neutrino mass term in Eq. (6) into $\mathcal{L}_{\rm Y}$. In this case one should also add the kinetic term of three right-handed neutrinos into $\mathcal{L}_{\rm F}$. It is straightforward to show that the conditions of CP invariance in the lepton sector turn out to be

$$
Y_\nu=Y^*_\nu\ ,\qquad Y_l=Y^*_l\ , \qquad (55)
$$

exactly in parallel with the quark sector. If an effective Majorana mass term is introduced into $\mathcal{L}_{\rm Y}$, as shown in Eq. (14), then the conditions of CP invariance in the lepton sector become

$$
M_{\rm L}=M^*_{\rm L}\ ,\quad Y_l=Y^*_l\ , \qquad (56)
$$

where $M_{\rm L}$ is the effective Majorana neutrino mass matrix. One may diagonalize both Y_ν (or $M_{\rm L}$) and Y_l to make them real and positive, but such a treatment will transfer CP violation from the Yukawa interactions to the weak charged-current interactions. Then lepton flavor mixing and CP violation are described by the 3×3 unitary matrix V given in Eq. (12), analogous to the 3×3 unitary matrix of quark flavor mixing and CP violation. In other words, the source of CP violation is the irremovable complex

phase(s) in the flavor mixing matrix of quarks or leptons. That is why we claim that CP violation stems from the coexistence of $\mathcal{L}_{\rm F}$ and $\mathcal{L}_{\rm Y}$ within the SM and, in most cases, beyond the SM.

It is worth reiterating that the process of spontaneous gauge symmetry breaking in the SM does not spontaneously violate CP. After the Higgs doublet H acquires its vacuum expectation value (i.e., $\phi^+ \to 0$ and $\phi^0 \to v/\sqrt{2}$ with v being real), we obtain three massive gauge bosons $W^\pm_\mu$ and Z_μ as well as one massless gauge boson A_μ. According to their relations with W^i_μ and B_μ, it is easy to find out the transformation properties of these physical fields under CP:

$$W^\pm_\mu \stackrel{\rm CP}{\longrightarrow} -W^{\mp\mu} \; , \quad Z_\mu \stackrel{\rm CP}{\longrightarrow} -Z^\mu \; , \quad A_\mu \stackrel{\rm CP}{\longrightarrow} -A^\mu \; , \tag{57}$$

with $\vec{x} \to -\vec{x}$ under P and CP for each field. In contrast, the neutral Higgs boson h is a CP-even particle. After spontaneous electroweak symmetry breaking, we are left with the quark mass matrices $M_{\rm u} = vY_{\rm u}/\sqrt{2}$ and $M_{\rm d} = vY_{\rm d}/\sqrt{2}$ or the lepton mass matrices $M_{\rm D} = vY_\nu/\sqrt{2}$ and $M_l = vY_l/\sqrt{2}$. The conditions of CP invariance given above can therefore be replaced with the corresponding mass matrices.

3 Electromagnetic Properties

3.1 Electromagnetic form factors

Although a neutrino does not possess any electric charge, it can have electromagnetic interactions via quantum loops. One may summarize such interactions by means of the following effective interaction term:

$$\mathcal{L}_{\rm EM} = \overline{\psi} \Gamma_\mu \psi A^\mu \equiv J_\mu(x) A^\mu(x) \; , \tag{58}$$

where the form of the electromagnetic current $J_\mu(x)$ is our present concern. Dirac and Majorana neutrinos couple to the photon in different ways, which are described by their respective electromagnetic form factors.

For an arbitrary Dirac particle (e.g., a Dirac neutrino), let us write down the matrix element of $J_\mu(x)$ between two one-particle states:

$$\langle \psi(p')|J_\mu(x)|\psi(p)\rangle = e^{-iqx}\langle \psi(p')|J_\mu(0)|\psi(p)\rangle = e^{-iqx}\overline{u}(\vec{p}')\Gamma_\mu(p,p')u(\vec{p}) \tag{59}$$

with $q = p - p'$. Because $J_\mu(x)$ is a Lorentz vector, the electromagnetic vertex function $\Gamma_\mu(p,p')$ must be a Lorentz vector too. The electromagnetic current conservation (or $U(1)_{\rm Q}$ gauge symmetry) requires $\partial^\mu J_\mu(x) = 0$, leading to

$$\langle \psi(p')|\partial^\mu J_\mu(x)|\psi(p)\rangle = (-iq^\mu)\, e^{-iqx}\overline{u}(\vec{p}')\Gamma_\mu(p,p')u(\vec{p}) = 0 \; . \tag{60}$$

Thus

$$q^\mu \overline{u}(\vec{p}')\Gamma_\mu(p,p')u(\vec{p}) = 0 \tag{61}$$

holds as one of the model-independent constraints on the form of $\Gamma_\mu(p,p')$. In addition, the Hermiticity of $J_\mu(x)$ or its matrix element implies

$$\begin{aligned} & e^{-iqx}\overline{u}(\vec{p}')\Gamma_\mu(p,p')u(\vec{p}) = e^{+iqx}\left[\overline{u}(\vec{p}')\Gamma_\mu(p,p')u(\vec{p})\right]^\dagger \\ & = e^{+iqx}\overline{u}(\vec{p})\left[\gamma_0\Gamma^\dagger_\mu(p,p')\gamma_0\right]u(\vec{p}') = e^{-iqx}\overline{u}(\vec{p}')\left[\gamma_0\Gamma^\dagger_\mu(p',p)\gamma_0\right]u(\vec{p}) \; , \end{aligned} \tag{62}$$

from which we immediately arrive at the second constraint on $\Gamma_\mu(p,p')$:

$$\Gamma_\mu(p,p') = \gamma_0\Gamma^\dagger_\mu(p',p)\gamma_0 \; . \tag{63}$$

Because of $p^2 = p'^2 = m^2$ with m being the fermion mass, we have $(p+p')^2 = 4m^2 - q^2$. Hence $\Gamma_\mu(p,p')$ depends only on the Lorentz-invariant quantity q^2.

A careful analysis of the Lorentz structure of $\overline{u}(\vec{p}')\Gamma_{\mu}(p,p')u(\vec{p})$, with the help of the Gordon-like identities and the constraints given above, shows that $\Gamma_{\mu}(p,p')$ may in general consist of four independent terms:

$$\Gamma_{\mu}(p,p') = f_{\rm Q}(q^2)\gamma_{\mu} + f_{\rm M}(q^2)i\sigma_{\mu\nu}q^{\nu} + f_{\rm E}(q^2)\sigma_{\mu\nu}q^{\nu}\gamma_5 + f_{\rm A}(q^2)\left(q^2\gamma_{\mu} - q_{\mu}\not{q}\right)\gamma_5 \,, \tag{64}$$

where $f_{\rm Q}(q^2)$, $f_{\rm M}(q^2)$, $f_{\rm E}(q^2)$ and $f_{\rm A}(q^2)$ are usually referred to as the charge, magnetic dipole, electric dipole and anapole form factors, respectively. In the non-relativistic limit of $\mathcal{L}_{\rm EM}$, it is easy to find that $f_{\rm Q}(0) = Q$ represents the electric charge of the particle, $f_{\rm M}(0) \equiv \mu$ denotes the magnetic dipole moment of the particle (i.e., $\mathcal{L}_{\rm EM}(f_{\rm M}) = -\mu\vec{\sigma}\cdot\vec{B}$ with $\vec{B}$ being the static magnetic field), $f_{\rm E}(0) \equiv \epsilon$ stands for the electric dipole moment of the particle (i.e., $\mathcal{L}_{\rm EM}(f_{\rm E}) = -\epsilon\vec{\sigma}\cdot\vec{E}$ with $\vec{E}$ being the static electric field), and $f_{\rm A}(0)$ corresponds to the Zeldovich anapole moment of the particle (i.e., $\mathcal{L}_{\rm EM}(f_{\rm A}) \propto f_{\rm A}(0)\vec{\sigma}\cdot[\nabla\times\vec{B} - \dot{\vec{E}}]$). One can observe that these form factors are not only Lorentz-invariant but also real (i.e., ${\rm Im}f_{\rm Q} = {\rm Im}f_{\rm M} = {\rm Im}f_{\rm E} = {\rm Im}f_{\rm A} = 0$). The latter is actually guaranteed by the Hermiticity condition in Eq. (62).

Given the form of Γ_{μ} in Eq. (64), it is straightforward to check the CP properties of $\mathcal{L}_{\rm EM}$ in Eq. (58). Note that the photon field transforms as $A^{\mu} \to -A_{\mu}$ under CP, and [4]

$$\begin{aligned}
\overline{\psi}\gamma_{\mu}\psi &\xrightarrow{\rm CP} -\overline{\psi}\gamma^{\mu}\psi \,, \\
\overline{\psi}\gamma_{\mu}\gamma_5\psi &\xrightarrow{\rm CP} -\overline{\psi}\gamma^{\mu}\gamma_5\psi \,, \\
\overline{\psi}\sigma_{\mu\nu}\psi &\xrightarrow{\rm CP} -\overline{\psi}\sigma^{\mu\nu}\psi \,, \\
\overline{\psi}\sigma_{\mu\nu}\gamma_5\psi &\xrightarrow{\rm CP} +\overline{\psi}\sigma^{\mu\nu}\gamma_5\psi \,.
\end{aligned} \tag{65}$$

Hence only the term proportional to $f_{\rm E}$ in $\mathcal{L}_{\rm EM}$ is CP-violating. If CP were conserved, then this term would vanish (i.e., $f_{\rm E} = 0$ would hold). Although there is no experimental hint at CP violation in the lepton sector, we expect that it should exist as in the quark sector. In any case, all four form factors are finite for a Dirac neutrino.

If neutrinos are massive Majorana particles, their electromagnetic properties will be rather different. The reason is simply that Majorana particles are their own antiparticles and thus can be described by using a smaller number of degrees of freedom. A free Majorana neutrino field ψ is by definition equal to its charge-conjugate field $\psi^c = C\overline{\psi}^T$ up to a global phase. Then

$$\overline{\psi}\Gamma_{\mu}\psi = \overline{\psi^c}\Gamma_{\mu}\psi^c = \psi^T\mathcal{C}\Gamma_{\mu}\mathcal{C}\overline{\psi}^T = \left(\psi^T\mathcal{C}\Gamma_{\mu}\mathcal{C}\overline{\psi}^T\right)^T = -\overline{\psi}\mathcal{C}^T\Gamma_{\mu}^T\mathcal{C}^T\psi \,, \tag{66}$$

from which one arrives at

$$\Gamma_{\mu} = -\mathcal{C}^T\Gamma_{\mu}^T\mathcal{C}^T = \mathcal{C}\Gamma_{\mu}^T\mathcal{C}^{-1} \,. \tag{67}$$

Substituting Eq. (64) into the right-hand side of Eq. (67) and taking account of $\mathcal{C}\gamma_{\mu}^T\mathcal{C}^{-1} = -\gamma_{\mu}$, $\mathcal{C}(\gamma_{\mu}\gamma_5)^T\mathcal{C}^{-1} = +\gamma_{\mu}\gamma_5$, $\mathcal{C}\sigma_{\mu\nu}^T\mathcal{C}^{-1} = -\sigma_{\mu\nu}$ and $\mathcal{C}(\sigma_{\mu\nu}\gamma_5)^T\mathcal{C}^{-1} = -\sigma_{\mu\nu}\gamma_5$, we obtain

$$\Gamma_{\mu}(p,p') = -f_{\rm Q}(q^2)\gamma_{\mu} - f_{\rm M}(q^2)i\sigma_{\mu\nu}q^{\nu} - f_{\rm E}(q^2)\sigma_{\mu\nu}q^{\nu}\gamma_5 + f_{\rm A}(q^2)\left(q^2\gamma_{\mu} - q_{\mu}\not{q}\right)\gamma_5 \,. \tag{68}$$

A comparison between Eqs. (64) and (68) yields

$$f_{\rm Q}(q^2) = f_{\rm M}(q^2) = f_{\rm E}(q^2) = 0 \,. \tag{69}$$

This result means that a Majorana neutrino only has the anapole form factor $f_{\rm A}(q^2)$.

[4] Taking account of $\mathcal{C}^{-1}\sigma_{\mu\nu}\mathcal{C} = -\sigma_{\mu\nu}^T$ and $\mathcal{C}^{-1}\gamma_5\mathcal{C} = \gamma_5^T$, one may easily prove that $\overline{\psi}\sigma_{\mu\nu}\gamma_5\psi$ is odd under both C and P. Thus $\overline{\psi}\sigma_{\mu\nu}\gamma_5\psi$ is CP-even.

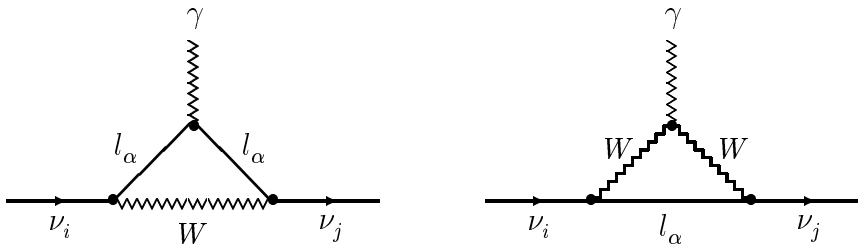


Fig. 1: One-loop Feynman diagrams contributing to the magnetic and electric dipole moments of massive Dirac neutrinos, where $\alpha = e, \mu, \tau$ and $i, j = 1, 2, 3$.

More generally, one may write out the matrix elements of the electromagnetic current $J_\mu(x)$ between two different states (i.e., the incoming and outgoing particles are different):

$$\langle \psi_j(p')|J_\mu(x)|\psi_i(p)\rangle = e^{-iqx}\overline{u_j}(\vec{p}')\Gamma^{ij}_\mu(p,p')u_i(\vec{p}) \; , \tag{70}$$

where $q = p - p'$ together with $p^2 = m_i^2$ and $p'^2 = m_j^2$ (for $i \neq j$). Here the electromagnetic vertex matrix $\Gamma_\mu(p,p')$ can be decomposed into the following Lorentz-invariant form in terms of four form factors:

$$\Gamma_\mu(p,p') = F_{\rm Q}(q^2)\left(q^2\gamma_\mu - q_\mu \not{q}\right) + F_{\rm M}(q^2) i\sigma_{\mu\nu}q^\nu + F_{\rm E}(q^2)\sigma_{\mu\nu}q^\nu\gamma_5 + F_{\rm A}(q^2)\left(q^2\gamma_\mu - q_\mu\not{q}\right)\gamma_5 \; , \tag{71}$$

where $F_{\rm Q}$, $F_{\rm M}$, $F_{\rm E}$ and $F_{\rm A}$ are all the 2×2 matrices in the space of neutrino mass eigenstates. The diagonal case (i.e., $i = j$) has been discussed above, from Eq. (59) to Eq. (69). In the off-diagonal case (i.e., $i \neq j$), the Hermiticity of $J_\mu(x)$ is no more a constraint on $\Gamma_\mu(p,p')$ for Dirac neutrinos because Eq. (62) only holds for $i = j$. It is now possible for Majorana neutrinos to have finite *transition* dipole moments, simply because Eqs. (66)—(69) do not hold when ψ_i and ψ_j represent different flavors.

We conclude that Dirac neutrinos may have both electric and magnetic dipole moments, while Majorana neutrinos have neither electric nor magnetic dipole moments. But massive Majorana neutrinos can have *transition* dipole moments which involve two different neutrino flavors in the initial and final states, so can massive Dirac neutrinos.

3.2 Magnetic and electric dipole moments

The magnetic and electric dipole moments of massive neutrinos, denoted as $\mu \equiv F_{\rm M}(0)$ and $\epsilon \equiv F_{\rm E}(0)$, are interesting in both theories and experiments because they are closely related to the dynamics of neutrino mass generation and to the characteristic of new physics.

Let us consider a minimal extension of the SM in which three right-handed neutrinos are introduced and lepton number conservation is required. In this case massive neutrinos are Dirac particles and their magnetic and electric dipole moments can be evaluated by calculating the Feynman diagrams in Fig. 1. Taking account of the smallness of both m_α^2/M_W^2 and m_i^2/M_W^2, where m_α (for $\alpha = e, \mu, \tau$) and m_i (for $i = 1, 2, 3$) stand respectively for the charged-lepton and neutrino masses, one obtains

$$\begin{aligned} \mu^{\rm D}_{ij} &= \frac{3eG_{\rm F}m_i}{32\sqrt{2}\pi^2}\left(1 + \frac{m_j}{m_i}\right) \times \sum_\alpha \left(2 - \frac{m_\alpha^2}{M_W^2}\right) V_{\alpha i}V^*_{\alpha j} \; , \\ \epsilon^{\rm D}_{ij} &= \frac{3eG_{\rm F}m_i}{32\sqrt{2}\pi^2}\left(1 - \frac{m_j}{m_i}\right) \times \sum_\alpha \left(2 - \frac{m_\alpha^2}{M_W^2}\right) V_{\alpha i}V^*_{\alpha j} \; , \end{aligned} \tag{72}$$

to an excellent degree of accuracy. Here $V_{\alpha i}$ and $V_{\alpha j}$ are the elements of the unitary lepton flavor mixing matrix V. Some discussions are in order.

(1) In the diagonal case (i.e., $i = j$), we are left with vanishing electric dipole moments (i.e., $\epsilon^{\rm D}_{ii} = 0$). The magnetic dipole moments $\mu^{\rm D}_{ii}$ are finite and proportional to the neutrino masses m_i (for $i = 1, 2, 3$):

$$\mu^{\rm D}_{ii} = \frac{3eG_{\rm F}m_i}{8\sqrt{2}\pi^2}\left(1 - \frac{1}{2}\sum_\alpha \frac{m^2_\alpha}{M^2_W}|V_{\alpha i}|^2\right) . \tag{73}$$

Hence a massless Dirac neutrino in the SM has no magnetic dipole moment. In the leading-order approximation, $\mu^{\rm D}_{ii}$ are independent of the strength of lepton flavor mixing and have tiny values

$$\mu^{\rm D}_{ii} \approx \frac{3eG_{\rm F}m_i}{8\sqrt{2}\pi^2} \approx 3\times 10^{-19}\left(\frac{m_i}{1\ {\rm eV}}\right)\mu_{\rm B} , \tag{74}$$

where $\mu_{\rm B} = e\hbar/(2m_e)$ is the Bohr magneton. Given $m_i \leq 1$ eV, the magnitude of $\mu^{\rm D}_{ii}$ is far below its present experimental upper bound ($<$ a few $\times\, 10^{-11}\mu_{\rm B}$).

(2) In the off-diagonal case (i.e., $i \neq j$), the unitarity of V allows us to simplify Eq. (72) to

$$\begin{aligned}
\mu^{\rm D}_{ij} &= -\frac{3eG_{\rm F}m_i}{32\sqrt{2}\pi^2}\left(1 + \frac{m_j}{m_i}\right)\sum_\alpha \frac{m^2_\alpha}{M^2_W}V_{\alpha i}V^*_{\alpha j} , \\
\epsilon^{\rm D}_{ij} &= -\frac{3eG_{\rm F}m_i}{32\sqrt{2}\pi^2}\left(1 - \frac{m_j}{m_i}\right)\sum_\alpha \frac{m^2_\alpha}{M^2_W}V_{\alpha i}V^*_{\alpha j} .
\end{aligned} \tag{75}$$

We see that the magnitudes of $\mu^{\rm D}_{ij}$ and $\epsilon^{\rm D}_{ij}$ (for $i \neq j$), compared with that of $\mu^{\rm D}_{ii}$, are further suppressed due to the smallness of m^2_α/M^2_W. Similar to the expression given in Eq. (74),

$$\begin{aligned}
\mu^{\rm D}_{ij} &\approx -4\times 10^{-23}\left(\frac{m_i + m_j}{1\ {\rm eV}}\right)\times\left(\sum_\alpha \frac{m^2_\alpha}{m^2_\tau}V_{\alpha i}V^*_{\alpha j}\right)\mu_{\rm B} , \\
\epsilon^{\rm D}_{ij} &\approx -4\times 10^{-23}\left(\frac{m_i - m_j}{1\ {\rm eV}}\right)\times\left(\sum_\alpha \frac{m^2_\alpha}{m^2_\tau}V_{\alpha i}V^*_{\alpha j}\right)\mu_{\rm B} ,
\end{aligned} \tag{76}$$

which can illustrate how small $\mu^{\rm D}_{ij}$ and $\epsilon^{\rm D}_{ij}$ are.

(3) Although Majorana neutrinos do not have intrinsic ($i = j$) magnetic and electric dipole moments, they may have finite transition ($i \neq j$) dipole moments. Because of the fact that Majorana neutrinos are their own antiparticles, their magnetic and electric dipole moments can also get contributions from two additional one-loop Feynman diagrams involving the charge-conjugate fields of ν_i, ν_j, l_α, $W^\pm$ and γ shown in Fig. 1 [5]. In this case one obtains

$$\begin{aligned}
\mu^{\rm M}_{ij} &= -\frac{3eG_{\rm F}i}{16\sqrt{2}\pi^2}\left(m_i + m_j\right)\times\sum_\alpha \frac{m^2_\alpha}{M^2_W}{\rm Im}\left(V_{\alpha i}V^*_{\alpha j}\right) , \\
\epsilon^{\rm M}_{ij} &= -\frac{3eG_{\rm F}}{16\sqrt{2}\pi^2}\left(m_i - m_j\right)\times\sum_\alpha \frac{m^2_\alpha}{M^2_W}{\rm Re}\left(V_{\alpha i}V^*_{\alpha j}\right) ,
\end{aligned} \tag{77}$$

where $m_i \neq m_j$ must hold. Comparing between Eqs. (75) and (77), we observe that the magnitudes of $\mu^{\rm M}_{ij}$ and $\epsilon^{\rm M}_{ij}$ are the same order as those of $\mu^{\rm D}_{ij}$ and $\epsilon^{\rm D}_{ij}$ in most cases, although the CP-violating phases hidden in $V_{\alpha i}V^*_{\alpha j}$ are possible to give rise to significant cancellations in some cases.

(4) The fact that μ_{ij} and ϵ_{ij} are proportional to m_i or m_j can be understood in the following way. Note that both tensor- and pseudotensor-like spinor bilinears are chirality-changing operators, which link the left-handed state to the right-handed one [6]:

$$\overline{\psi}\sigma_{\mu\nu}\psi = \overline{\psi_{\rm L}}\sigma_{\mu\nu}\psi_{\rm R} + {\rm h.c.} ,$$

[5] Here we confine ourselves to a simple extension of the SM with three known neutrinos to be massive Majorana particles.

[6] That is why both magnetic and electric dipole moments must vanish for a Weyl neutrino, because it is massless and does not possess the right-handed component.

$$\overline{\psi}\sigma_{\mu\nu}\gamma_5\psi = \overline{\psi_{\rm L}}\sigma_{\mu\nu}\gamma_5\psi_{\rm R} - {\rm h.c.}\; . \tag{78}$$

Note also that the same relations hold when ψ is replaced by its charge-conjugate field ψ^c for Majorana neutrinos. Because $(\nu_i)_{\rm R}$ and $(\nu_j)_{\rm R}$ do not have any interactions with $W^\pm$ in Fig. 1, it seems that only $(\nu_i)_{\rm L}$ and $(\nu_j)_{\rm L}$ are flowing along the external fermion lines. To obtain a chirality-changing contribution from the effective (one-loop) electromagnetic vertex, one has to put a mass insertion on one of the external legs in the Feynman diagrams. As a result, the magnetic and electric dipole moments must involve m_i and m_j, the masses of ν_i and ν_j neutrinos.

(5) Is the magnetic or electric dipole moment of a neutrino always proportional to its mass? The answer is negative if new physics beyond the $SU(2)_{\rm L} \times U(1)_{\rm Y}$ gauge theory is involved. For instance, a new term proportional to the charged-lepton mass can contribute to the magnetic dipole moment of a massive Dirac neutrino in the $SU(2)_{\rm L} \times SU(2)_{\rm R} \times U(1)_{\rm Y}$ model with broken left-right symmetry. Depending on the details of this model, such a term might cancel or exceed the term proportional to the neutrino mass in the expression of the magnetic dipole moment.

Finite magnetic and electric dipole moments of massive neutrinos may produce a variety of new processes beyond the SM. For example, (a) radiative neutrino decays $\nu_i \to \nu_j + \gamma$ can happen, so can the Cherenkov radiation of neutrinos in an external electromagnetic field; (b) the elastic neutrino-electron or neutrino-nucleon scattering can be mediated by the magnetic and electric dipole moments; (c) the phenomenon of precession of the neutrino spin can occur in an external magnetic field; (d) the photon (or plasmon) can decay into a neutrino-antineutrino pair in a plasma (i.e., $\gamma^* \to \nu\overline{\nu}$). Of course, non-vanishing electromagnetic dipole moments contribute to neutrino masses too.

3.3 Radiative neutrino decays

If the electromagnetic moments of a massive neutrino ν_i are finite, it can decay into a lighter neutrino ν_j and a photon γ. The Lorentz-invariant vertex matrix of this $\nu_i \to \nu_j + \gamma$ process is in general described by $\Gamma_\mu(p, p')$ in Eq. (71). Because $q^2 = 0$ and $q_\mu \varepsilon^\mu = 0$ hold for a real photon γ, where ε^μ represents the photon polarization, the form of $\Gamma_\mu(p, p')$ can be simplified to

$$\Gamma_\mu(p, p') = \left[iF_{\rm M}(0) + F_{\rm E}(0)\gamma_5 \right] \sigma_{\mu\nu} q^\nu \; . \tag{79}$$

By definition, $F^{ij}_{\rm M}(0) \equiv \mu_{ij}$ and $F^{ij}_{\rm E}(0) \equiv \epsilon_{ij}$ are just the magnetic and electric transition dipole moments between ν_i and ν_j neutrinos. Given the transition matrix element $\overline{u_j}(\vec{p}\,')\Gamma^{ij}_\mu(p, p')u_i(\vec{p})$, it is straightforward to calculate the decay rate. In the rest frame of the decaying neutrino ν_i,

$$\Gamma_{\nu_i \to \nu_j + \gamma} = \frac{\left(m_i^2 - m_j^2 \right)^3}{8\pi m_i^3} \left(\left| \mu_{ij} \right|^2 + \left| \epsilon_{ij} \right|^2 \right) \; . \tag{80}$$

This result is valid for both Dirac and Majorana neutrinos.

In the $SU(2)_{\rm L} \times U(1)_{\rm Y}$ gauge theory with three massive Dirac (or Majorana) neutrinos, the radiative decay $\nu_i \to \nu_j + \gamma$ is mediated by the one-loop Feynman diagrams (and their charge-conjugate diagrams) shown in Fig. 1. The explicit expressions of μ_{ij} and ϵ_{ij} have been given in Eq. (75) for Dirac neutrinos and in Eq. (77) for Majorana neutrinos. Hence

$$\Gamma^{\rm (D)}_{\nu_i \to \nu_j + \gamma} = \frac{\left(m_i^2 - m_j^2 \right)^3}{8\pi m_i^3} \left(\left| \mu^{\rm D}_{ij} \right|^2 + \left| \epsilon^{\rm D}_{ij} \right|^2 \right) = \frac{9\alpha G_{\rm F}^2 m_i^5}{2^{11}\pi^4} \left(1 - \frac{m_j^2}{m_i^2} \right)^3 \left(1 + \frac{m_j^2}{m_i^2} \right) \times \left| \sum_\alpha \frac{m_\alpha^2}{M_W^2} V_{\alpha i} V^*_{\alpha j} \right|^2 \; , \tag{81}$$

for Dirac neutrinos; or

$$\Gamma^{(\rm M)}_{\nu_i\to\nu_j+\gamma} = \frac{\left(m_i^2 - m_j^2\right)^3}{8\pi m_i^3}\left(\left|\mu^{\rm M}_{ij}\right|^2 + \left|\epsilon^{\rm M}_{ij}\right|^2\right) = \frac{9\alpha G_{\rm F}^2 m_i^5}{2^{10}\pi^4}\left(1-\frac{m_j^2}{m_i^2}\right)^3\left\{\left(1+\frac{m_j}{m_i}\right)^2 \times \left[\sum_\alpha \frac{m_\alpha^2}{M_W^2}{\rm Im}\left(V_{\alpha i}V^*_{\alpha j}\right)\right]^2 + \left(1-\frac{m_j}{m_i}\right)^2\left[\sum_\alpha \frac{m_\alpha^2}{M_W^2}{\rm Re}\left(V_{\alpha i}V^*_{\alpha j}\right)\right]^2\right\} , \quad (82)$$

for Majorana neutrinos, where $\alpha = e^2/(4\pi)$ denotes the electromagnetic fine-structure constant.

To compare $\Gamma_{\nu_i\to\nu_j+\gamma}$ with the experimental data in a simpler way, one may define an effective magnetic dipole moment

$$\mu_{\rm eff} \equiv \sqrt{\left|\mu_{ij}\right|^2 + \left|\epsilon_{ij}\right|^2} \ . \quad (83)$$

Eq. (80) can then be expressed as

$$\Gamma_{\nu_i\to\nu_j+\gamma} = 5.3 \times \left(1-\frac{m_j^2}{m_i^2}\right)^3 \left(\frac{m_i}{1\ {\rm eV}}\right)^3 \times \left(\frac{\mu_{\rm eff}}{\mu_{\rm B}}\right)^2 {\rm s}^{-1} \ . \quad (84)$$

Although $\mu_{\rm eff}$ is extremely small in some simple extensions of the SM, it could be sufficiently large in some more complicated or exotic scenarios beyond the SM, such as a class of extra-dimension models. Experimentally, radiative decays of massive neutrinos can be constrained by seeing no emission of the photons from solar ν_e and reactor $\overline{\nu}_e$ fluxes. Much stronger constraints on $\mu_{\rm eff}$ can be obtained from the Supernova 1987A limit on the neutrino decay and from the astrophysical limit on distortions of the cosmic microwave background (CMB) radiation. A brief summary of these limits is

$$\frac{\mu_{\rm eff}}{\mu_{\rm B}} < \begin{cases} 0.9\times 10^{-1}\left(\dfrac{\rm eV}{m_\nu}\right)^2 & \text{Reactor} \\ 0.5\times 10^{-5}\left(\dfrac{\rm eV}{m_\nu}\right)^2 & \text{Sun} \\ 1.5\times 10^{-8}\left(\dfrac{\rm eV}{m_\nu}\right)^2 & \text{SN 1987A} \\ 1.0\times 10^{-11}\left(\dfrac{\rm eV}{m_\nu}\right)^{9/4} & \text{CMB} \end{cases}$$

where m_ν denotes the effective mass of the decaying neutrino (i.e., $m_\nu = m_i$).

3.4 Electromagnetic ν_e-e scattering

In practice, the most sensitive way of probing the electromagnetic dipole moments of a massive neutrino is to measure the cross section of elastic neutrino-electron (or antineutrino-electron) scattering, which can be expressed as a sum of the contribution from the SM (σ_0) and that from the electromagnetic dipole moments of massive neutrinos (σ_μ):

$$\frac{{\rm d}\sigma}{{\rm d}T} = \frac{{\rm d}\sigma_0}{{\rm d}T} + \frac{{\rm d}\sigma_\mu}{{\rm d}T} \ , \quad (85)$$

where $T = E_e - m_e$ denotes the kinetic energy of the recoil electron in this process. We have

$$\frac{{\rm d}\sigma_0}{{\rm d}T} = \frac{G_{\rm F}^2 m_e}{2\pi}\left[g_+^2 + g_-^2\left(1-\frac{T}{E_\nu}\right)^2 - g_+ g_- \frac{m_e T}{E_\nu^2}\right] \quad (86)$$

for neutrino-electron scattering, where $g_+ = 2\sin^2\theta_{\rm w} + 1$ for ν_e, $g_+ = 2\sin^2\theta_{\rm w} - 1$ for ν_μ and ν_τ, and $g_- = 2\sin^2\theta_{\rm w}$ for all flavors. Note that Eq. (86) is also valid for antineutrino-electron scattering if one simply exchanges the positions of g_+ and g_-. On the other hand,

$$\frac{{\rm d}\sigma_\mu}{{\rm d}T} = \frac{\alpha^2\pi}{m_e^2}\left(\frac{1}{T} - \frac{1}{E_\nu}\right)\left(\frac{\mu_\nu}{\mu_{\rm B}}\right)^2 \tag{87}$$

with $\mu_\nu^2 \equiv |\mu_{ii}^{\rm D}|^2 + |\epsilon_{ii}^{\rm D}|^2$ (for $i = 1$, 2 or 3), which holds for both neutrinos and antineutrinos. In obtaining Eqs. (86) and (87) one has assumed the scattered neutrino to be a Dirac particle and omitted the effects of finite neutrino masses and flavor mixing (i.e., $\nu_e = \nu_1$, $\nu_\mu = \nu_2$ and $\nu_\tau = \nu_3$ have been taken). Hence there is no interference between the contributions coming from the SM and electromagnetic dipole moments — the latter leads to a helicity flip of the neutrino but the former is always helicity-conserving. While an interference term will appear if one takes account of neutrino masses and flavor mixing, its magnitude linearly depends on the neutrino masses and thus is strongly suppressed in comparison with the pure weak and electromagnetic terms. So the incoherent sum of ${\rm d}\sigma_0/{\rm d}T$ and ${\rm d}\sigma_\mu/{\rm d}T$ in Eq. (85) is actually an excellent approximation of ${\rm d}\sigma/{\rm d}T$.

It is obvious that the two terms of ${\rm d}\sigma/{\rm d}T$ depend on the kinetic energy of the recoil electron in quite different ways. In particular, ${\rm d}\sigma_\mu/{\rm d}T$ grows rapidly with decreasing values of T. Hence a measurement of smaller T can probe smaller μ_ν in this kind of experiments. The magnitude of ${\rm d}\sigma_\mu/{\rm d}T$ becomes larger than that of ${\rm d}\sigma_0/{\rm d}T$ if the condition

$$T \leq \frac{\alpha^2\pi^2}{G_{\rm F}^2 m_e^3}\left(\frac{\mu_\nu}{\mu_{\rm B}}\right)^2 \approx 3\times 10^{22}\left(\frac{\mu_\nu}{\mu_{\rm B}}\right)^2 {\rm keV} \tag{88}$$

is roughly satisfied, as one can easily see from Eqs. (86) and (87). No distortion of the recoil electron energy spectrum of $\nu_\alpha e^-$ or $\overline{\nu}_\alpha e^-$ scattering (for $\alpha = e, \mu, \tau$) has so far been observed in any direct laboratory experiments, and thus only the upper bounds on μ_ν can be derived. For instance, an analysis of the T-spectrum in the Super-Kamiokande experiment yields $\mu_\nu < 1.1\times 10^{-10}\mu_{\rm B}$. More stringent bounds on μ_ν can hopefully be achieved in the future.

In view of current experimental data on neutrino oscillations, we know that neutrinos are actually massive. Hence the effects of finite neutrino masses and flavor mixing should be taken into account in calculating the cross section of elastic neutrino-electron or antineutrino-electron scattering. Here let us illustrate how the neutrino oscillation may affect the weak and electromagnetic terms of elastic $\overline{\nu}_e e^-$ scattering in a reactor experiment, where the antineutrinos are produced from the beta decay of fission products and detected by their elastic scattering with electrons in a detector. The antineutrino state created in this beta decay (via $W^- \to e^- + \overline{\nu}_e$) at the reactor is a superposition of three antineutrino mass eigenstates:

$$|\overline{\nu}_e(0)\rangle = \sum_{j=1}^{3} V_{ej}|\overline{\nu}_j\rangle \; . \tag{89}$$

Such a $\overline{\nu}_e$ beam propagates over the distance L to the detector,

$$|\overline{\nu}_e(L)\rangle = \sum_{j=1}^{3} e^{iq_jL}V_{ej}|\overline{\nu}_j\rangle \; , \tag{90}$$

in which $q_j = \sqrt{E_\nu^2 - m_j^2}$ is the momentum of ν_j with E_ν being the beam energy and m_j being the mass of ν_j. After taking account of the effect of neutrino oscillations, one obtains the differential cross section of elastic antineutrino-electron scattering as follows:

$$\frac{{\rm d}\sigma'}{{\rm d}T} = \frac{{\rm d}\sigma'_0}{{\rm d}T} + \frac{{\rm d}\sigma'_\mu}{{\rm d}T} \; , \tag{91}$$

where

$$\frac{\mathrm{d}\sigma_0'}{\mathrm{d}T} = \frac{G_{\rm F}^2 m_e}{2\pi} \left\{ g_-^2 + \left(g_- - 1\right)^2 \left(1 - \frac{T}{E_\nu}\right)^2 - g_- \left(g_- - 1\right) \frac{m_e T}{E_\nu^2} + 2 g_- \left| \sum_{j=1}^{3} e^{i q_j L} |V_{ej}|^2 \right|^2 \left[2 \left(1 - \frac{T}{E_\nu}\right)^2 - \frac{m_e T}{E_\nu^2} \right] \right\} \tag{92}$$

with $g_- = 2 \sin^2 \theta_{\rm w}$ for $\overline{\nu}_e$, and

$$\frac{\mathrm{d}\sigma_\mu'}{\mathrm{d}T} = \frac{\alpha^2 \pi}{m_e^2} \sum_{k=1}^{3} \left| \sum_{j=1}^{3} e^{i q_j L} V_{ej} \frac{\epsilon_{jk} + i \mu_{jk}}{\mu_{\rm B}} \right|^2 \times \left(\frac{1}{T} - \frac{1}{E_\nu} \right) \tag{93}$$

with μ_{jk} and ϵ_{jk} being the magnetic and electric transition dipole moments between ν_j and ν_k neutrinos as defined in Eq. (79). Because different neutrino mass eigenstates are in principle distinguishable in the electromagnetic $\overline{\nu}_e e^-$ scattering, their contributions to the total cross section are incoherent. Eq. (93) shows that it is in general difficult to determine or constrain the magnitudes of μ_{jk} and ϵ_{jk} (for $j, k = 1, 2, 3$) from a single measurement.

4 Lepton Flavor Mixing and CP Violation

Regardless of the dynamical origin of tiny neutrino masses [7], we may discuss lepton flavor mixing by taking account of the effective mass terms of charged leptons and Majorana neutrinos at low energies [8],

$$-\mathcal{L}'_{\rm lepton} = \overline{(e\ \mu\ \tau)_{\rm L}}\, M_l \begin{pmatrix} e \\ \mu \\ \tau \end{pmatrix}_{\rm R} + \frac{1}{2} \overline{\left(\nu_e\ \nu_\mu\ \nu_\tau\right)_{\rm L}}\, M_\nu \begin{pmatrix} \nu_e^c \\ \nu_\mu^c \\ \nu_\tau^c \end{pmatrix}_{\rm R} + {\rm h.c.}\ . \tag{94}$$

The phenomenon of lepton flavor mixing arises from a mismatch between the diagonalizations of M_l and M_ν in an arbitrary flavor basis: $V_l^\dagger M_l U_l = {\rm Diag}\{m_e, \mu_\mu, m_\tau\}$ and $V_\nu^\dagger M_\nu V_\nu^* = {\rm Diag}\{m_1, m_2, m_3\}$, where V_l, U_l and V_ν are the 3×3 unitary matrices. In the basis of mass eigenstates, it is the unitary matrix $V = V_l^\dagger V_\nu$ that will appear in the weak charged-current interactions in Eq. (12). Although the basis of $M_l = {\rm Diag}\{m_e, m_\mu, m_\tau\}$ with $V_l = \mathbf{1}$ and $V = V_\nu$ is often chosen in neutrino phenomenology, one should keep in mind that both the charged-lepton and neutrino sectors may in general contribute to lepton flavor mixing. In other words, both V_l and V_ν are not fully physical, and only their product $V = V_l^\dagger V_\nu$ is a physical description of lepton flavor mixing and CP violation at low energies.

4.1 Parametrizations of V

Flavor mixing among n different lepton families can be described by an $n \times n$ unitary matrix V, whose number of independent parameters relies on the nature of neutrinos. If neutrinos are Dirac particles, one may make use of $n(n-1)/2$ rotation angles and $(n-1)(n-2)/2$ phase angles to parametrize V. If neutrinos are Majorana particles, however, a full parametrization of V needs $n(n-1)/2$ rotation angles and the same number of phase angles [9]. The flavor mixing between charged leptons and Dirac

[7] For simplicity, here we do not consider possible non-unitarity of the 3×3 neutrino mixing matrix because its effects are either absent or very small.

[8] As for Dirac neutrinos, the corresponding mass term is the same as that given in Eq. (7). In this case the neutrino mass matrix M_ν is in general not symmetric and can be diagonalized by means of the transformation $V_\nu^\dagger M_\nu U_\nu = {\rm Diag}\{m_1, m_2, m_3\}$, where both V_ν and U_ν are unitary.

[9] No matter whether neutrinos are Dirac or Majorana particles, the $n \times n$ unitary flavor mixing matrix has $(n-1)^2(n-2)^2/4$ Jarlskog invariants of CP violation defined as $\mathcal{J}_{\alpha\beta}^{ij} \equiv {\rm Im}\left(V_{\alpha i} V_{\beta j} V_{\alpha j}^* V_{\beta i}^*\right)$.

neutrinos is completely analogous to that of quarks, for which a number of different parametrizations have been proposed and classified in the literature. Here we classify all possible parametrizations for the flavor mixing between charged leptons and Majorana neutrinos with $n = 3$. Regardless of the freedom of phase reassignments, we find that there are nine structurally different parametrizations for the 3×3 lepton flavor mixing matrix V.

The 3×3 lepton flavor mixing matrix V, which is often called the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix, can be expressed as a product of three unitary matrices O_1, O_2 and O_3. They correspond to simple rotations in the complex (1,2), (2,3) and (3,1) planes:

$$O_1 = \begin{pmatrix} c_1 e^{i\alpha_1} & s_1 e^{-i\beta_1} & 0 \\ -s_1 e^{i\beta_1} & c_1 e^{-i\alpha_1} & 0 \\ 0 & 0 & e^{i\gamma_1} \end{pmatrix} ,$$

$$O_2 = \begin{pmatrix} e^{i\gamma_2} & 0 & 0 \\ 0 & c_2 e^{i\alpha_2} & s_2 e^{-i\beta_2} \\ 0 & -s_2 e^{i\beta_2} & c_2 e^{-i\alpha_2} \end{pmatrix} ,$$

$$O_3 = \begin{pmatrix} c_3 e^{i\alpha_3} & 0 & s_3 e^{-i\beta_3} \\ 0 & e^{i\gamma_3} & 0 \\ -s_3 e^{i\beta_3} & 0 & c_3 e^{-i\alpha_3} \end{pmatrix} , \tag{95}$$

where $s_i \equiv \sin\theta_i$ and $c_i \equiv \cos\theta_i$ (for $i = 1, 2, 3$). Obviously $O_i O_i^\dagger = O_i^\dagger O_i = \mathbf{1}$ holds, and any two rotation matrices do not commute with each other. We find twelve different ways to arrange the product of O_1, O_2 and O_3, which can cover the whole 3×3 space and provide a full description of V. Explicitly, six of the twelve different combinations of O_i belong to the type

$$V = O_i(\theta_i, \alpha_i, \beta_i, \gamma_i) \otimes O_j(\theta_j, \alpha_j, \beta_j, \gamma_j) \otimes O_i(\theta_i', \alpha_i', \beta_i', \gamma_i') \tag{96}$$

with $i \neq j$, where the complex rotation matrix O_i occurs twice; and the other six belong to the type

$$V = O_i(\theta_i, \alpha_i, \beta_i, \gamma_i) \otimes O_j(\theta_j, \alpha_j, \beta_j, \gamma_j) \otimes O_k(\theta_k, \alpha_k, \beta_k, \gamma_k) \tag{97}$$

with $i \neq j \neq k$, in which the rotations take place in three different complex planes. The products $O_i O_j O_i$ and $O_i O_k O_i$ (for $i \neq k$) in Eq. (97) are correlated with each other, if the relevant phase parameters are switched off. Hence only nine of the twelve parametrizations, three from Eq. (96) and six from Eq. (97), are structurally different.

In each parametrization of V, there apparently exist nine phase parameters. Some of them or their combinations can be absorbed by redefining the relevant phases of charged-lepton and neutrino fields. If neutrinos are Dirac particles, V contains only a single irremovable CP-violating phase δ. If neutrinos are Majorana particles, however, there is no freedom to rearrange the relative phases of three Majorana neutrino fields. Hence V may in general contain three irremovable CP-violating phases in the Majorana case (δ and two Majorana phases). Both CP- and T-violating effects in neutrino oscillations depend only upon the Dirac-like phase δ.

Different parametrizations of V are mathematically equivalent, so adopting any of them does not directly point to physical significance. But it is very likely that one particular parametrization is more useful and transparent than the others in studying the neutrino phenomenology and (or) exploring the underlying dynamics responsible for lepton mass generation and CP violation. Here we highlight two particular parametrizations of the PMNS matrix V. The first one is the so-called "standard" parametrization advocated by the Particle Data Group:

$$V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{pmatrix} \begin{pmatrix} c_{13} & 0 & s_{13} e^{-i\delta} \\ 0 & 1 & 0 \\ -s_{13} e^{i\delta} & 0 & c_{13} \end{pmatrix} \begin{pmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{pmatrix} P' , \tag{98}$$

where $c_{ij} \equiv \cos\theta_{ij}$ and $s_{ij} \equiv \sin\theta_{ij}$ (for $ij = 12, 13, 23$) together with the Majorana phase matrix $P' = \mathrm{Diag}\{e^{i\rho}, e^{i\sigma}, 1\}$. Without loss of generality, the three mixing angles $(\theta_{12}, \theta_{13}, \theta_{23})$ can all be arranged to lie in the first quadrant. Arbitrary values between 0 and 2π are allowed for three CP-violating phases (δ, ρ, σ). A remarkable merit of this parametrization is that its three mixing angles are approximately equivalent to the mixing angles of solar (θ_{12}), atmospheric (θ_{23}) and CHOOZ reactor (θ_{13}) neutrino oscillation experiments. Another useful parametrization is the Fritzsch-Xing (FX) parametrization proposed originally for quark mixing and later for lepton mixing:

$$V = \begin{pmatrix} c_l & s_l & 0 \\ -s_l & c_l & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} e^{-i\phi} & 0 & 0 \\ 0 & c & s \\ 0 & -s & c \end{pmatrix} \begin{pmatrix} c_\nu & -s_\nu & 0 \\ s_\nu & c_\nu & 0 \\ 0 & 0 & 1 \end{pmatrix} P' \; , \tag{99}$$

where $c_{l,\nu} \equiv \cos\theta_{l,\nu}$, $s_{l,\nu} \equiv \sin\theta_{l,\nu}$, $c \equiv \cos\theta$, $s \equiv \sin\theta$, and P' is a diagonal phase matrix containing two nontrivial CP-violating phases. Although the form of V in Eq. (99) is apparently different from that in Eq. (98), their corresponding flavor mixing angles $(\theta_l, \theta_\nu, \theta)$ and $(\theta_{12}, \theta_{13}, \theta_{23})$ have quite similar meanings in interpreting the experimental data on neutrino oscillations. In the limit $\theta_l = \theta_{13} = 0$, one easily arrives at $\theta_\nu = \theta_{12}$ and $\theta = \theta_{23}$. As a natural consequence of very small θ_l, three mixing angles of the FX parametrization can also be related to those of solar (θ_ν), atmospheric (θ) and CHOOZ reactor $(\theta_l \sin\theta)$ neutrino oscillation experiments in the leading-order approximation. A striking merit of this parametrization is that its six parameters have very simple renormalization-group equations when they run from a superhigh-energy scale to the electroweak scale or vice versa.

4.2 Democratic or tri-bimaximal mixing?

Current neutrino oscillation data indicate the essential feature of lepton flavor mixing: two mixing angles are quite large ($\theta_{12} \sim 34^\circ$ and $\theta_{23} \sim 45^\circ$) while the third one is very small ($\theta_{13} < 10^\circ$). Such a flavor mixing pattern is far beyond the original imagination of most people because it is rather different from the well-known quark mixing pattern ($\vartheta_{12} \approx 14.5^\circ$, $\vartheta_{23} \approx 2.6^\circ$, $\vartheta_{13} \approx 0.23^\circ$ and $\delta = 76.5^\circ$) described by the same parametrization of the Cabibbo-Kobayashi-Maskawa (CKM) matrix. To understand this difference, a number of constant lepton mixing patterns have been proposed as the starting point of model building. Possible flavor symmetries and their spontaneous or explicit breaking mechanisms hidden in those constant patterns might finally help us pin down the dynamics responsible for lepton mass generation and flavor mixing. To illustrate, let us first comment on the "democratic" neutrino mixing pattern and then pay more attention to the "tri-bimaximal" neutrino mixing pattern.

The "democratic" lepton flavor mixing pattern

$$U_0 = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{\sqrt{2}}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix} \tag{100}$$

was originally obtained by Fritzsch and Xing as the leading term of the 3×3 lepton mixing matrix from the breaking of flavor democracy or $S(3)_{\rm L} \times S(3)_{\rm R}$ symmetry of the charged-lepton mass matrix in the basis where the Majorana neutrino mass matrix is diagonal and possesses the $S(3)$ symmetry. Its naive predictions $\theta_{12} = 45^\circ$ and $\theta_{23} \approx 54.7^\circ$ are no more favored today, but they may receive proper corrections from the symmetry-breaking perturbations so as to fit current neutrino oscillation data.

Today's most popular constant pattern of neutrino mixing is the "tri-bimaximal" mixing matrix:

$$V_0 = \begin{pmatrix} \frac{\sqrt{2}}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 \\ \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \end{pmatrix} \tag{101}$$

which looks like a twisted form of the democratic mixing pattern with the same entries. Its strange name comes from the fact that this flavor mixing pattern is actually a product of the "tri-maximal" mixing matrix and a "bi-maximal" mixing matrix:

$$V_0' = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{\omega}{\sqrt{3}} & \frac{\omega^2}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{\omega^2}{\sqrt{3}} & \frac{\omega}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix} = P V_0 P' \; , \tag{102}$$

where $\omega = e^{i2\pi/3}$ denotes the complex cube-root of unity (i.e., $\omega^3 = 1$), and $P = {\rm Diag}\{1, \omega, \omega^2\}$ and $P' = {\rm Diag}\{1, 1, i\}$ are two diagonal phase matrices. V_0 or V_0' predicts $\theta_{12} = \arctan(1/\sqrt{2}) \approx 35.3^\circ$, $\theta_{13} = 0^\circ$ and $\theta_{23} = 45^\circ$, consistent quite well with current neutrino oscillation data. Because the entries of U_0 or V_0 are all formed from small integers (0, 1, 2 and 3) and their square roots, it is often suggestive of certain discrete flavor symmetries in the language of group theories. That is why the democratic or tri-bimaximal neutrino mixing pattern can serve as a good starting point of model building based on a variety of flavor symmetries, such as Z_2, Z_3, S_3, S_4, A_4, D_4, D_5, Q_4, Q_6, $\Delta(27)$ and $\Sigma(81)$. In particular, a lot of interest has been paid to the derivation of V_0 with the help of the non-Abelian discrete A_4 symmetry.

Note that the democratic mixing matrix U_0 and the tri-bimaximal mixing matrix V_0 are related with each other via the following transformation:

$$V_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_0 & -\sin\theta_0 \\ 0 & \sin\theta_0 & \cos\theta_0 \end{pmatrix} U_0 \begin{pmatrix} \cos\theta_0 & -\sin\theta_0 & 0 \\ \sin\theta_0 & \cos\theta_0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \; , \tag{103}$$

where $\theta_0 = \arctan(\sqrt{2} - 1)^2 \approx 9.7^\circ$. This angle is actually a measure of the difference between the mixing angles of U_0 and V_0 (namely, $45^\circ - 35.3^\circ = 54.7^\circ - 45^\circ = 9.7^\circ$). In this sense, we argue that it is worthwhile to explore possible flavor symmetries behind both V_0 and U_0 so as to build realistic models for neutrino mass generation and lepton flavor mixing.

Let us remark that a specific constant mixing pattern should be regarded as the leading-order approximation of the "true" lepton flavor mixing matrix, whose mixing angles should in general depend on both the ratios of charged-lepton masses and those of neutrino masses. We may at least make the following naive speculation about how to phenomenologically understand the observed pattern of lepton flavor mixing:

- Large values of θ_{12} and θ_{23} could arise from a weak hierarchy or a near degeneracy of the neutrino mass spectrum, because the strong hierarchy of charged-lepton masses implies that m_e/m_μ and m_μ/m_τ at the electroweak scale are unlikely to contribute to θ_{12} and θ_{23} in a dominant way.
- Special values of θ_{12} and θ_{23} might stem from an underlying flavor symmetry of the charged-lepton mass matrix or the neutrino mass matrix. Then the contributions of lepton mass ratios to flavor mixing angles, due to flavor symmetry breaking, are expected to serve as perturbative corrections to U_0 or V_0, or another constant mixing pattern.
- Vanishing or small θ_{13} could be a natural consequence of the explicit textures of lepton mass matrices. It might also be related to the flavor symmetry which gives rise to sizable θ_{12} and θ_{23} (e.g., in U_0 or V_0).
- Small corrections to a constant flavor mixing pattern may also result from the renormalization-group running effects of leptons and quarks, e.g., from a superhigh-energy scale to low energies or vice versa.

There are too many possibilities of linking the observed pattern of lepton flavor mixing to a certain flavor symmetry, and none of them is unique from the theoretical point of view. In this sense, flavor symmetries should not be regarded as a perfect guiding principle of model building.

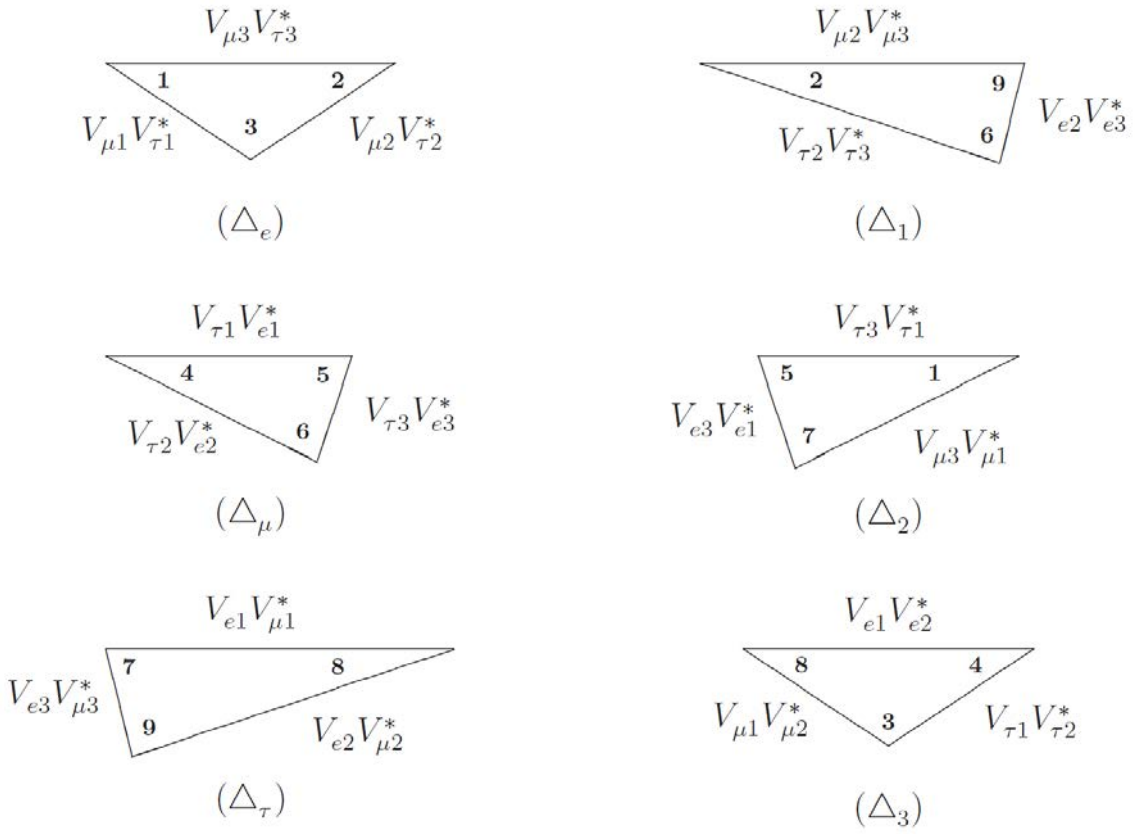


Fig. 2: Unitarity triangles of the 3×3 PMNS matrix in the complex plane. Each triangle is named by the index that does not manifest in its three sides.

4.3 Leptonic unitarity triangles

In the basis where the flavor eigenstates of charged leptons are identified with their mass eigenstates, the PMNS matrix V relates the neutrino mass eigenstates (ν_1, ν_2, ν_3) to the neutrino flavor eigenstates $(\nu_e, \nu_\mu, \nu_\tau)$:

$$\begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix} = \begin{pmatrix} V_{e1} & V_{e2} & V_{e3} \\ V_{\mu1} & V_{\mu2} & V_{\mu3} \\ V_{\tau1} & V_{\tau2} & V_{\tau3} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} \; . \tag{104}$$

The unitarity of V represents two sets of normalization and orthogonality conditions:

$$\sum_i \left(V_{\alpha i} V^*_{\beta i}\right) = \delta_{\alpha\beta} \; , \quad \sum_\alpha \left(V_{\alpha i} V^*_{\alpha j}\right) = \delta_{ij} \; , \tag{105}$$

where Greek and Latin subscripts run over (e, μ, τ) and $(1, 2, 3)$, respectively. In the complex plane the six orthogonality relations in Eq. (105) define six triangles $(\triangle_e, \triangle_\mu, \triangle_\tau)$ and $(\triangle_1, \triangle_2, \triangle_3)$ shown in Fig. 2, the so-called unitarity triangles. These six triangles have eighteen different sides and nine different inner (or outer) angles. But the unitarity of V requires that all six triangles have the same area amounting to $\mathcal{J}/2$, where $\mathcal{J}$ is the Jarlskog invariant of CP violation defined through

$$\mathrm{Im}\left(V_{\alpha i} V_{\beta j} V^*_{\alpha j} V^*_{\beta i}\right) = \mathcal{J} \sum_\gamma \epsilon_{\alpha\beta\gamma} \sum_k \epsilon_{ijk} \; . \tag{106}$$

One has $\mathcal{J} = c_{12} s_{12} c^2_{13} s_{13} c_{23} s^2_{23} \sin\delta$ in the standard parametrization of V as well as $\mathcal{J} = c_l s_l c_\nu s_\nu c s^2 \sin\phi$ in the FX parametrization of V. No matter whether neutrinos are Dirac or Majorana particles, the strength of CP or T violation in neutrino oscillations depends only upon $\mathcal{J}$.

To show why the areas of six unitarity triangles are identical with one another, let us take triangles $\triangle_\tau$ and $\triangle_3$ for example. They correspond to the orthogonality relations

$$\begin{aligned} V_{e1} V^*_{\mu 1} + V_{e2} V^*_{\mu 2} + V_{e3} V^*_{\mu 3} &= 0 \; , \\ V_{e1} V^*_{e2} + V_{\mu 1} V^*_{\mu 2} + V_{\tau 1} V^*_{\tau 2} &= 0 \; . \end{aligned} \tag{107}$$

Multiplying these two equations by $V_{\mu 2} V^*_{e2}$ and $V_{\mu 2} V^*_{\mu 1}$ respectively, we arrive at two rescaled triangles which share the side

$$V_{e1} V_{\mu 2} V^*_{e2} V^*_{\mu 1} = -|V_{e2} V_{\mu 2}|^2 - V_{e3} V_{\mu 2} V^*_{e2} V^*_{\mu 3} = -|V_{\mu 1} V_{\mu 2}|^2 - V_{\mu 2} V_{\tau 1} V^*_{\mu 1} V^*_{\tau 2} \; . \tag{108}$$

This result is consistent with the definition of $\mathcal{J}$ in Eq. (106); i.e., $\mathrm{Im}(V_{e1} V_{\mu 2} V^*_{e2} V^*_{\mu 1}) = \mathcal{J}$ and $\mathrm{Im}(V_{e3} V_{\mu 2} V^*_{e2} V^*_{\mu 3}) = \mathrm{Im}(V_{\mu 2} V_{\tau 1} V^*_{\mu 1} V^*_{\tau 2}) = -\mathcal{J}$. The latter simultaneously implies that the areas of $\triangle_\tau$ and $\triangle_3$ are equal to $\mathcal{J}/2$. One may analogously prove that all the six unitarity triangles have the same area $\mathcal{J}/2$. If CP or T were an exact symmetry, $\mathcal{J} = 0$ would hold and those unitarity triangles would collapse into lines in the complex plane. Note that the shape and area of each unitarity triangle are irrelevant to the nature of neutrinos; i.e., they are the same for Dirac and Majorana neutrinos.

Because of $V^*_{e1} V_{\mu 1} + V^*_{e2} V_{\mu 2} = -V^*_{e3} V_{\mu 3}$ or equivalently $|V_{e1} V^*_{\mu 1} + V_{e2} V^*_{\mu 2}|^2 = |V_{e3} V^*_{\mu 3}|^2$, it is easy to obtain

$$2\mathrm{Re}\left(V_{e1} V_{\mu 2} V^*_{e2} V^*_{\mu 1}\right) = |V_{e3}|^2 |V_{\mu 3}|^2 - |V_{e1}|^2 |V_{\mu 1}|^2 - |V_{e2}|^2 |V_{\mu 2}|^2 \; . \tag{109}$$

Combining $V_{e1} V_{\mu 2} V^*_{e2} V^*_{\mu 1} = \mathrm{Re}(V_{e1} V_{\mu 2} V^*_{e2} V^*_{\mu 1}) + i\mathcal{J}$ with Eq. (109) leads us to the result

$$\begin{aligned} \mathcal{J}^2 &= |V_{e1}|^2 |V_{\mu 2}|^2 |V_{e2}|^2 |V_{\mu 1}|^2 - \frac{1}{4} \left(|V_{e3}|^2 |V_{\mu 3}|^2 - |V_{e1}|^2 |V_{\mu 1}|^2 - |V_{e2}|^2 |V_{\mu 2}|^2\right)^2 \\ &= |V_{e1}|^2 |V_{\mu 2}|^2 |V_{e2}|^2 |V_{\mu 1}|^2 - \frac{1}{4} \left(1 + |V_{e1}|^2 |V_{\mu 2}|^2 + |V_{e2}|^2 |V_{\mu 1}|^2 \right. \\ &\left. \hspace{10em} - |V_{e1}|^2 - |V_{\mu 2}|^2 - |V_{e2}|^2 - |V_{\mu 1}|^2\right)^2 \; . \end{aligned} \tag{110}$$

As a straightforward generalization of Eq. (110), $\mathcal{J}^2$ can be expressed in terms of the moduli of any four independent matrix elements of V:

$$\begin{aligned} \mathcal{J}^2 = |V_{\alpha i}|^2 |V_{\beta j}|^2 |V_{\alpha j}|^2 |V_{\beta i}|^2 - \frac{1}{4} \left(1 + |V_{\alpha i}|^2 |V_{\beta j}|^2 + |V_{\alpha j}|^2 |V_{\beta i}|^2 \right. \\ \left. - |V_{\alpha i}|^2 - |V_{\beta j}|^2 - |V_{\alpha j}|^2 - |V_{\beta i}|^2\right)^2 \; , \end{aligned} \tag{111}$$

in which $\alpha \neq \beta$ running over (e, μ, τ) and $i \neq j$ running over $(1, 2, 3)$. The implication of this result is very obvious: the information about leptonic CP violation can in principle be extracted from the measured moduli of the neutrino mixing matrix elements.

As a consequence of the unitarity of V, two interesting relations can be derived from the normalization conditions in Eq. (105):

$$\begin{aligned} |V_{e2}|^2 - |V_{\mu 1}|^2 = |V_{\mu 3}|^2 - |V_{\tau 2}|^2 = |V_{\tau 1}|^2 - |V_{e3}|^2 \equiv \Delta_{\rm L} \; , \\ |V_{e2}|^2 - |V_{\mu 3}|^2 = |V_{\mu 1}|^2 - |V_{\tau 2}|^2 = |V_{\tau 3}|^2 - |V_{e1}|^2 \equiv \Delta_{\rm R} \; . \end{aligned} \tag{112}$$

The off-diagonal asymmetries $\Delta_{\rm L}$ and $\Delta_{\rm R}$ characterize the geometrical structure of V about its V_{e1}-$V_{\mu 2}$-$V_{\tau 3}$ and V_{e3}-$V_{\mu 2}$-$V_{\tau 1}$ axes, respectively. For instance, $\Delta_{\rm L} = 1/6$ and $\Delta_{\rm R} = -1/6$ hold for the tri-bimaximal neutrino mixing pattern V_0. If $\Delta_{\rm L} = 0$ (or $\Delta_{\rm R} = 0$) held, V would be symmetric about the V_{e1}-$V_{\mu 2}$-$V_{\tau 3}$ (or V_{e3}-$V_{\mu 2}$-$V_{\tau 1}$) axis. Geometrically this would correspond to the congruence between two unitarity triangles; i.e.,

$$\Delta_{\rm L} = 0 : \; \triangle_e \cong \triangle_1 \; , \triangle_\mu \cong \triangle_2 \; , \triangle_\tau \cong \triangle_3 \; ;$$

Table 2: Some important discoveries in the developments of flavor physics.

	Discoveries of lepton flavors, quark flavors and CP violation
1897	electron (Thomson, 1897)
1919	proton (up and down quarks) (Rutherford, 1919)
1932	neutron (up and down quarks) (Chadwick, 1932)
1933	positron (Anderson, 1933)
1936	muon (Neddermeyer and Anderson, 1937)
1947	Kaon (strange quark) (Rochester and Butler, 1947)
1956	electron antineutrino (Cowan *et al.*, 1956)
1962	muon neutrino (Danby *et al.*, 1962)
1964	CP violation in s-quark decays (Christenson *et al.*, 1964)
1974	charm quark (Aubert *et al.*, 1974; Abrams *et al.*, 1974)
1975	tau (Perl *et al.*, 1975)
1977	bottom quark (Herb *et al.*, 1977)
1995	top quark (Abe *et al.*, 1995; Abachi *et al.*, 1995)
2000	tau neutrino (Kodama *et al.*, 2000)
2001	CP violation in b-quark decays (Aubert *et al.*, 2001; Abe *et al.*, 2001)

$$\Delta_{\rm R} = 0 : \ \triangle_e \cong \triangle_3 \ , \triangle_\mu \cong \triangle_2 \ , \triangle_\tau \cong \triangle_1 \ . \tag{113}$$

Indeed the counterpart of $\Delta_{\rm L}$ in the quark sector is only of $\mathcal{O}(10^{-5})$; i.e., the CKM matrix is almost symmetric about its V_{ud}-V_{cs}-V_{tb} axis. An exactly symmetric flavor mixing matrix might hint at an underlying flavor symmetry, from which some deeper understanding of the fermion mass texture could be achieved.

4.4 Flavor problems in particle physics

In the subatomic world the fundamental building blocks of matter have twelve flavors: six quarks and six leptons (and their antiparticles). Table 2 is a brief list of some important discoveries in flavor physics, which can partly give people a ball-park feeling of a century of developments in particle physics. The SM of electromagnetic and weak interactions contain thirteen free parameters in its lepton and quark sectors: three charged-lepton masses, six quark masses, three quark flavor mixing angles and one CP-violating phase. If three known neutrinos are massive Majorana particles, one has to introduce nine free parameters to describe their flavor properties: three neutrino masses, three lepton flavor mixing angles and three CP-violating phases. Thus an effective theory of electroweak interactions at low energies totally consists of twenty-two flavor parameters which can only be determined from experiments. Why is the number of degrees of freedom so big in the flavor sector? What is the fundamental physics behind these parameters? Such puzzles constitute the flavor problems in particle physics.

Current experimental data on neutrino oscillations can only tell us $m_1 < m_2$. It remains unknown whether m_3 is larger than m_2 (normal hierarchy) or smaller than m_1 (inverted hierarchy). The possibility $m_1 \approx m_2 \approx m_3$ (near degeneracy) cannot be excluded at present. In contrast, three families of charged fermions have very strong mass hierarchies:

$$\begin{aligned} &\frac{m_e}{m_\mu} \sim \frac{m_u}{m_c} \sim \frac{m_c}{m_t} \sim \lambda^4 \ , \\ &\frac{m_\mu}{m_\tau} \sim \frac{m_d}{m_s} \sim \frac{m_s}{m_b} \sim \lambda^2 \ , \end{aligned} \tag{114}$$

where $\lambda \equiv \sin\theta_{\rm C} \approx 0.22$ with $\theta_{\rm C}$ being the Cabibbo angle of quark flavor mixing. In the standard

parametrization of the CKM matrix, three quark mixing angles exhibit an impressive hierarchy:

$$\vartheta_{12} \sim \lambda \; , \quad \vartheta_{23} \sim \lambda^2 \; , \quad \vartheta_{13} \sim \lambda^4 \; . \tag{115}$$

These two kinds of hierarchies might intrinsically be related to each other, because the flavor mixing angles actually measure a mismatch between the mass and flavor eigenstates of up- and down-type quarks. For example, the relations $\vartheta_{12} \approx \sqrt{m_d/m_s}$, $\vartheta_{23} \approx \sqrt{m_d/m_b}$ and $\vartheta_{13} \approx \sqrt{m_u/m_t}$ are compatible with Eqs. (114) and (115). They can be derived from a specific pattern of up- and down-type quark mass matrices with five texture zeros. On the other hand, it seems quite difficult to find a simple way of linking two large lepton flavor mixing angles $\theta_{12} \sim \pi/6$ and $\theta_{23} \sim \pi/4$ to small m_e/m_μ and m_μ/m_τ. One might ascribe the largeness of θ_{12} and θ_{23} to a very weak hierarchy of three neutrino masses and the smallness of θ_{13} to the strong mass hierarchy in the charged-lepton sector. There are of course many possibilities of model building to understand the observed lepton flavor mixing pattern, but none of them has experimentally and theoretically been justified.

Among a number of concrete flavor puzzles that are currently facing us, the following three are particularly intriguing.

– The pole masses of three charged leptons satisfy the equality

$$\frac{m_e + m_\mu + m_\tau}{\left(\sqrt{m_e} + \sqrt{m_\mu} + \sqrt{m_\tau}\right)^2} = \frac{2}{3} \tag{116}$$

to an amazingly good degree of accuracy — its error bar is only of $\mathcal{O}(10^{-5})$.

– There are two quark-lepton "complementarity" relations in flavor mixing:

$$\theta_{12} + \vartheta_{12} \approx \theta_{23} + \vartheta_{23} \approx \frac{\pi}{4} \; , \tag{117}$$

which are compatible with the present experimental data.

– Two unitarity triangles of the CKM matrix, defined by the orthogonality conditions $V_{ud}V^*_{ub} + V_{cd}V^*_{cb} + V_{td}V^*_{tb} = 0$ and $V_{tb}V^*_{ub} + V_{ts}V^*_{us} + V_{td}V^*_{ud} = 0$, are almost the right triangles. Namely, the common inner angle of these two triangles satisfies

$$\alpha \equiv \arg\left(-\frac{V_{ud}V^*_{ub}}{V_{td}V^*_{tb}}\right) \approx \frac{\pi}{2} \; , \tag{118}$$

indicated by current experimental data on quark mixing and CP violation.

Such special numerical relations might just be accidental. One or two of them might also be possible to result from a certain (underlying) flavor symmetry.

5 Running of Neutrino Mass Parameters

5.1 One-loop RGEs

The spirit of seesaw mechanisms is to attribute the small masses of three known neutrinos to the existence of some heavy degrees of freedom, such as the $SU(2)_{\rm L}$ gauge-singlet fermions, the $SU(2)_{\rm L}$ gauge-triplet scalars or the $SU(2)_{\rm L}$ gauge-triplet fermions. All of them point to the unique dimension-5 Weinberg operator in an effective theory after the corresponding heavy particles are integrated out:

$$\frac{\mathcal{L}_{\rm d=5}}{\Lambda} = \frac{1}{2}\kappa_{\alpha\beta}\overline{\ell_{\alpha\rm L}}\tilde{H}\tilde{H}^T\ell^c_{\beta\rm L} + {\rm h.c.} \; , \tag{119}$$

where Λ is the cutoff scale, $\ell_{\rm L}$ denotes the left-handed lepton doublet, $\tilde{H} \equiv i\sigma_2 H^*$ with H being the SM Higgs doublet, and κ stands for the effective neutrino coupling matrix. After spontaneous gauge

symmetry breaking, $\tilde{H}$ gains its vacuum expectation value $\langle \tilde{H} \rangle = v/\sqrt{2}$ with $v \approx 246$ GeV. We are then left with the effective Majorana mass matrix $M_\nu = \kappa v^2/2$ for three light neutrinos from Eq. (119). If the dimension-5 Weinberg operator is obtained in the framework of the minimal supersymmetric standard model (MSSM), one will be left with $M_\nu = \kappa (v \sin\beta)^2/2$, where $\tan\beta$ denotes the ratio of the vacuum expectation values of two MSSM Higgs doublets.

Eq. (119) or its supersymmetric counterpart can provide a simple but generic way of generating tiny neutrino masses. There are a number of interesting possibilities of building renormalizable gauge models to realize the effective Weinberg mass operator, either radiatively or at the tree level. The latter case is just associated with the well-known seesaw mechanisms to be discussed in section 6. Here we assume that $\mathcal{L}_{\rm d=5}/\Lambda$ arises from an underlying seesaw model, whose lightest heavy particle has a mass of $\mathcal{O}(\Lambda)$. In other words, Λ characterizes the seesaw scale. Above Λ there may exist one or more energy thresholds corresponding to the masses of heavier seesaw particles. Below Λ the energy dependence of the effective neutrino coupling matrix κ is described by its renormalization-group equation (RGE). The evolution of κ from Λ down to the electroweak scale is formally independent of any details of the relevant seesaw model from which κ is derived.

At the one-loop level κ obeys the RGE

$$16\pi^2 \frac{{\rm d}\kappa}{{\rm d}t} = \alpha_\kappa \kappa + C_\kappa \left[(Y_l Y_l^\dagger)\kappa + \kappa (Y_l Y_l^\dagger)^T \right] \tag{120}$$

where $t \equiv \ln(\mu/\Lambda)$ with μ being an arbitrary renormalization scale between the electroweak scale and the seesaw scale, and Y_l is the charged-lepton Yukawa coupling matrix. The RGE of Y_l and those of $Y_{\rm u}$ (up-type quarks) and $Y_{\rm d}$ (down-type quarks) are given by

$$\begin{aligned}
16\pi^2 \frac{{\rm d}Y_l}{{\rm d}t} &= \left[\alpha_l + C_l^l (Y_l Y_l^\dagger) \right] Y_l \; , \\
16\pi^2 \frac{{\rm d}Y_{\rm u}}{{\rm d}t} &= \left[\alpha_{\rm u} + C_{\rm u}^{\rm u} (Y_{\rm u} Y_{\rm u}^\dagger) + C_{\rm u}^{\rm d} (Y_{\rm d} Y_{\rm d}^\dagger) \right] Y_{\rm u} \; , \\
16\pi^2 \frac{{\rm d}Y_{\rm d}}{{\rm d}t} &= \left[\alpha_{\rm d} + C_{\rm d}^{\rm u} (Y_{\rm u} Y_{\rm u}^\dagger) + C_{\rm d}^{\rm d} (Y_{\rm d} Y_{\rm d}^\dagger) \right] Y_{\rm d} \; .
\end{aligned} \tag{121}$$

In the framework of the SM we have

$$\begin{aligned}
C_\kappa &= C_{\rm u}^{\rm d} = C_{\rm d}^{\rm u} = -\frac{3}{2} \; , \\
C_l^l &= C_{\rm u}^{\rm u} = C_{\rm d}^{\rm d} = +\frac{3}{2} \; ,
\end{aligned} \tag{122}$$

and

$$\begin{aligned}
\alpha_\kappa &= -3g_2^2 + \lambda + 2{\rm Tr}\left[3(Y_{\rm u} Y_{\rm u}^\dagger) + 3(Y_{\rm d} Y_{\rm d}^\dagger) + (Y_l Y_l^\dagger) \right] \; , \\
\alpha_l &= -\frac{9}{4} g_1^2 - \frac{9}{4} g_2^2 + {\rm Tr}\left[3(Y_{\rm u} Y_{\rm u}^\dagger) + 3(Y_{\rm d} Y_{\rm d}^\dagger) + (Y_l Y_l^\dagger) \right] \; , \\
\alpha_{\rm u} &= -\frac{17}{20} g_1^2 - \frac{9}{4} g_2^2 - 8 g_3^2 + {\rm Tr}\left[3(Y_{\rm u} Y_{\rm u}^\dagger) + 3(Y_{\rm d} Y_{\rm d}^\dagger) + (Y_l Y_l^\dagger) \right] \; , \\
\alpha_{\rm d} &= -\frac{1}{4} g_1^2 - \frac{9}{4} g_2^2 - 8 g_3^2 + {\rm Tr}\left[3(Y_{\rm u} Y_{\rm u}^\dagger) + 3(Y_{\rm d} Y_{\rm d}^\dagger) + (Y_l Y_l^\dagger) \right] \; ;
\end{aligned} \tag{123}$$

and in the framework of the MSSM we have

$$\begin{aligned}
C_\kappa &= C_{\rm u}^{\rm d} = C_{\rm d}^{\rm u} = +1 \; , \\
C_l^l &= C_{\rm u}^{\rm u} = C_{\rm d}^{\rm d} = +3 \; ,
\end{aligned} \tag{124}$$

and

$$\alpha_\kappa = -\frac{6}{5} g_1^2 - 6 g_2^2 + 6{\rm Tr}(Y_{\rm u} Y_{\rm u}^\dagger) \; ,$$

$$\alpha_l = -\frac{9}{5} g_1^2 - 3 g_2^2 + {\rm Tr}\left[3(Y_{\rm d} Y_{\rm d}^\dagger) + (Y_l Y_l^\dagger)\right] \; ,$$
$$\alpha_{\rm u} = -\frac{13}{15} g_1^2 - 3 g_2^2 - \frac{16}{3} g_3^2 + 3{\rm Tr}(Y_{\rm u} Y_{\rm u}^\dagger) \; ,$$
$$\alpha_{\rm d} = -\frac{7}{15} g_1^2 - 3 g_2^2 - \frac{16}{3} g_3^2 + {\rm Tr}\left[3(Y_{\rm d} Y_{\rm d}^\dagger) + (Y_l Y_l^\dagger)\right] \; . \tag{125}$$

Here g_1, g_2 and g_3 are the gauge couplings and satisfy their RGEs

$$16\pi^2 \frac{{\rm d}g_i}{{\rm d}t} = b_i g_i^3 \; , \tag{126}$$

where $(b_1, b_2, b_3) = (41/10, -19/6, -7)$ in the SM or $(33/5, 1, -3)$ in the MSSM. In addition, λ is the Higgs self-coupling parameter of the SM and obeys the RGE

$$\begin{aligned} 16\pi^2 \frac{{\rm d}\lambda}{{\rm d}t} = {} & 6\lambda^2 - 3\lambda \left(\frac{3}{5} g_1^2 + 3 g_2^2\right) + \frac{3}{2} \left(\frac{3}{5} g_1^2 + g_2^2\right)^2 + 3 g_2^4 \\ & + 4\lambda {\rm Tr}\left[3(Y_{\rm u} Y_{\rm u}^\dagger) + 3(Y_{\rm d} Y_{\rm d}^\dagger) + (Y_l Y_l^\dagger)\right] \\ & - 8 {\rm Tr}\left[3(Y_{\rm u} Y_{\rm u}^\dagger)^2 + 3(Y_{\rm d} Y_{\rm d}^\dagger)^2 + (Y_l Y_l^\dagger)^2\right] \; . \end{aligned} \tag{127}$$

The relation between λ and the Higgs mass M_h is given by $\lambda = M_h^2/(2v^2)$, where $v \approx 246$ GeV is the vacuum expectation value of the Higgs field.

The above RGEs allow us to evaluate the running behavior of κ together with those of Y_l, $Y_{\rm u}$ and $Y_{\rm d}$, from the seesaw scale to the electroweak scale or vice versa. We shall examine the evolution of neutrino masses, lepton flavor mixing angles and CP-violating phases in the following.

5.2 Running neutrino mass parameters

Without loss of any generality, we choose the flavor basis where Y_l is diagonal: $Y_l = D_l \equiv {\rm Diag}\{y_e, y_\mu, y_\tau\}$ with y_α being the eigenvalues of Y_l. In this case the effective Majorana neutrino coupling matrix κ can be diagonalized by the PMNS matrix V; i.e., $V^\dagger \kappa V^* = \widehat{\kappa} \equiv {\rm Diag}\{\kappa_1, \kappa_2, \kappa_3\}$ with κ_i being the eigenvalues of κ. Then

$$\frac{{\rm d}\kappa}{{\rm d}t} = \dot{V} \widehat{\kappa} V^T + V \dot{\widehat{\kappa}} V^T + V \widehat{\kappa} \dot{V}^T = \frac{1}{16\pi^2} \left[\alpha_\kappa V \widehat{\kappa} V^T + C_\kappa \left(D_l^2 V \widehat{\kappa} V^T + V \widehat{\kappa} V^T D_l^2\right)\right] \; , \tag{128}$$

with the help of Eq. (120). After a definition of the Hermitian matrix $S \equiv V^\dagger D_l^2 V$ and the anti-Hermitian matrix $T \equiv V^\dagger \dot{V}$, Eq. (128) leads to

$$\dot{\widehat{\kappa}} = \frac{1}{16\pi^2} \left[\alpha_\kappa \widehat{\kappa} + C_\kappa (S \widehat{\kappa} + \widehat{\kappa} S^*)\right] - T \widehat{\kappa} + \widehat{\kappa} T^* . \tag{129}$$

Because $\widehat{\kappa}$ is by definition diagonal and real, the left- and right-hand sides of Eq. (129) must be diagonal and real. We can therefore arrive at

$$\dot{\kappa}_i = \frac{1}{16\pi^2} \left(\alpha_\kappa + 2 C_\kappa {\rm Re} S_{ii}\right) \kappa_i \; , \tag{130}$$

together with ${\rm Im} T_{ii} = {\rm Re} T_{ii} = {\rm Im} S_{ii} = 0$ (for $i = 1, 2, 3$). As the off-diagonal parts of Eq. (129) are vanishing, we have

$$T_{ij} \kappa_j - \kappa_i T_{ij}^* = \frac{C_\kappa}{16\pi^2} \left(S_{ij} \kappa_j + \kappa_i S_{ij}^*\right) \tag{131}$$

with $i \neq j$. Therefore,

$${\rm Re} T_{ij} = -\frac{C_\kappa}{16\pi^2} \frac{\kappa_i + \kappa_j}{\kappa_i - \kappa_j} {\rm Re} S_{ij} \; ,$$

$$\mathrm{Im}T_{ij} = -\frac{C_\kappa}{16\pi^2}\frac{\kappa_i - \kappa_j}{\kappa_i + \kappa_j}\mathrm{Im}S_{ij} \; . \tag{132}$$

Due to $\dot{V} = VT$, Eq. (132) actually governs the evolution of V with energies.

We proceed to define $V \equiv PUP'$, in which $P \equiv \mathrm{Diag}\{e^{i\phi_e}, e^{i\phi_\mu}, e^{i\phi_\tau}\}$, $P' \equiv \mathrm{Diag}\{e^{i\rho}, e^{i\sigma}, 1\}$, and U is the CKM-like matrix containing three neutrino mixing angles and one CP-violating phase. Although P does not have any physical meaning, its phases have their own RGEs. In contrast, P' serves for the Majorana phase matrix. We find

$$T' \equiv P'TP'^\dagger = P'V^\dagger\dot{V}P'^\dagger = \dot{P}'P'^\dagger + U^\dagger\dot{U} + U^\dagger P^\dagger\dot{P}U \; , \tag{133}$$

from which we can obtain six independent constraint equations:

$$\begin{aligned}
T'_{11} &= i\dot{\rho} + \sum_\alpha \left[U^*_{\alpha 1}\dot{U}_{\alpha 1} + iU_{\alpha 1}\dot{\phi}_\alpha\right] \; , \\
T'_{22} &= i\dot{\sigma} + \sum_\alpha \left[U^*_{\alpha 2}\dot{U}_{\alpha 2} + iU_{\alpha 2}\dot{\phi}_\alpha\right] \; , \\
T'_{33} &= \sum_\alpha \left[U^*_{\alpha 3}\dot{U}_{\alpha 3} + iU_{\alpha 3}\dot{\phi}_\alpha\right] \; ; \\
T'_{12} &= \sum_\alpha \left[U^*_{\alpha 1}\dot{U}_{\alpha 2} + iU_{\alpha 2}\dot{\phi}_\alpha\right] \; , \\
T'_{13} &= \sum_\alpha \left[U^*_{\alpha 1}\dot{U}_{\alpha 3} + iU_{\alpha 3}\dot{\phi}_\alpha\right] \; , \\
T'_{23} &= \sum_\alpha \left[U^*_{\alpha 2}\dot{U}_{\alpha 3} + iU_{\alpha 3}\dot{\phi}_\alpha\right] \; ,
\end{aligned} \tag{134}$$

where α runs over e, μ and τ. Note that $T_{ii} = 0$ holds and T_{ij} is given by Eq. (132). In view of $y_e \ll y_\mu \ll y_\tau$, we take $D_l^2 \approx \mathrm{Diag}\{0, 0, y_\tau^2\}$ as an excellent approximation. Then S_{ij}, T_{ij} and T'_{ij} can all be expressed in terms of y_τ^2 and the parameters of U and P'. After a straightforward calculation, we obtain the explicit expressions of Eqs. (130) and (134) as follows:

$$\dot{\kappa}_i = \frac{\kappa_i}{16\pi^2}\left(\alpha_\kappa + 2C_\kappa y_\tau^2 |U_{\tau i}|^2\right) \; , \tag{135}$$

and

$$\begin{aligned}
&\sum_\alpha \left[U^*_{\alpha 1}\left(i\dot{U}_{\alpha 1} - U_{\alpha 1}\dot{\phi}_\alpha\right)\right] = \dot{\rho} \; , \\
&\sum_\alpha \left[U^*_{\alpha 2}\left(i\dot{U}_{\alpha 2} - U_{\alpha 2}\dot{\phi}_\alpha\right)\right] = \dot{\sigma} \; , \\
&\sum_\alpha \left[U^*_{\alpha 3}\left(i\dot{U}_{\alpha 3} - U_{\alpha 3}\dot{\phi}_\alpha\right)\right] = 0 \; , \\
&\sum_\alpha \left[U^*_{\alpha 1}\left(\dot{U}_{\alpha 2} + iU_{\alpha 2}\dot{\phi}_\alpha\right)\right] = -\frac{C_\kappa y_\tau^2}{16\pi^2}e^{i(\rho-\sigma)}\left[\zeta_{12}^{-1}\mathrm{Re}\left(U^*_{\tau 1}U_{\tau 2}e^{i(\sigma-\rho)}\right) + i\zeta_{12}\mathrm{Im}\left(U^*_{\tau 1}U_{\tau 2}e^{i(\sigma-\rho)}\right)\right] \\
&\sum_\alpha \left[U^*_{\alpha 1}\left(\dot{U}_{\alpha 3} + iU_{\alpha 3}\dot{\phi}_\alpha\right)\right] = -\frac{C_\kappa y_\tau^2}{16\pi^2}e^{i\rho}\left[\zeta_{13}^{-1}\mathrm{Re}\left(U^*_{\tau 1}U_{\tau 3}e^{-i\rho}\right) + i\zeta_{13}\mathrm{Im}\left(U^*_{\tau 1}U_{\tau 3}e^{-i\rho}\right)\right] \; , \\
&\sum_\alpha \left[U^*_{\alpha 2}\left(\dot{U}_{\alpha 3} + iU_{\alpha 3}\dot{\phi}_\alpha\right)\right] = -\frac{C_\kappa y_\tau^2}{16\pi^2}e^{i\sigma}\left[\zeta_{23}^{-1}\mathrm{Re}\left(U^*_{\tau 2}U_{\tau 3}e^{-i\sigma}\right) + i\zeta_{23}\mathrm{Im}\left(U^*_{\tau 2}U_{\tau 3}e^{-i\sigma}\right)\right] \; ,
\end{aligned} \tag{13}$$

where $\zeta_{ij} \equiv (\kappa_i - \kappa_j)/(\kappa_i + \kappa_j)$ with $i \neq j$. One can see that those y_τ^2-associated terms only consist of the matrix elements $U_{\tau i}$ (for $i = 1, 2, 3$). If a parametrization of U assures $U_{\tau i}$ to be as simple as

possible, the resultant RGEs of neutrino mixing angles and CP-violating phases will be very concise. We find that the FX parametrization advocated in Eq. (99) with

$$U = \begin{pmatrix} s_l s_\nu c + c_l c_\nu e^{-i\phi} & s_l c_\nu c - c_l s_\nu e^{-i\phi} & s_l s \\ c_l s_\nu c - s_l c_\nu e^{-i\phi} & c_l c_\nu c + s_l s_\nu e^{-i\phi} & c_l s \\ -s_\nu s & -c_\nu s & c \end{pmatrix}$$

accords with the above observation, while the "standard" parametrization in Eq. (98) does not. That is why the RGEs of neutrino mixing angles and CP-violating phases in the standard parametrization are rather complicated.

Here we take the FX form of U to derive the RGEs of neutrino mass and mixing parameters. Combining Eqs. (135), (136) and the FX form of U, we arrive at

$$\begin{aligned} \dot{\kappa}_1 &= \frac{\kappa_1}{16\pi^2}\left(\alpha_\kappa + 2C_\kappa y_\tau^2 s_\nu^2 s^2\right) \; , \\ \dot{\kappa}_2 &= \frac{\kappa_2}{16\pi^2}\left(\alpha_\kappa + 2C_\kappa y_\tau^2 c_\nu^2 s^2\right) \; , \\ \dot{\kappa}_3 &= \frac{\kappa_3}{16\pi^2}\left(\alpha_\kappa + 2C_\kappa y_\tau^2 c^2\right) \; , \end{aligned} \tag{137}$$

where $\alpha_\kappa \approx -3g_2^2 + 6y_t^2 + \lambda$ (SM) or $\alpha_\kappa \approx -1.2g_1^2 - 6g_2^2 + 6y_t^2$ (MSSM); and

$$\begin{aligned} \dot{\theta}_l &= \frac{C_\kappa y_\tau^2}{16\pi^2}\, c_\nu s_\nu c \left[\zeta_{13}^{-1} c_\rho c_{(\rho-\phi)} + \zeta_{13} s_\rho s_{(\rho-\phi)} - \zeta_{23}^{-1} c_\sigma c_{(\sigma-\phi)} - \zeta_{23} s_\sigma s_{(\sigma-\phi)}\right] \; , \\ \dot{\theta}_\nu &= \frac{C_\kappa y_\tau^2}{16\pi^2}\, c_\nu s_\nu \left[s^2\left(\zeta_{12}^{-1} c_{(\sigma-\rho)}^2 + \zeta_{12} s_{(\sigma-\rho)}^2\right) + c^2\left(\zeta_{13}^{-1} c_\rho^2 + \zeta_{13} s_\rho^2\right) - c^2\left(\zeta_{23}^{-1} c_\sigma^2 + \zeta_{23} s_\sigma^2\right)\right] \; , \\ \dot{\theta} &= \frac{C_\kappa y_\tau^2}{16\pi^2}\, cs \left[s_\nu^2\left(\zeta_{13}^{-1} c_\rho^2 + \zeta_{13} s_\rho^2\right) + c_\nu^2\left(\zeta_{23}^{-1} c_\sigma^2 + \zeta_{23} s_\sigma^2\right)\right] \; ; \end{aligned} \tag{138}$$

as well as

$$\begin{aligned} \dot{\rho} &= \frac{C_\kappa y_\tau^2}{16\pi^2}\left[\widehat{\zeta}_{12} c_\nu^2 s^2 c_{(\sigma-\rho)} s_{(\sigma-\rho)} + \widehat{\zeta}_{13}\left(s_\nu^2 s^2 - c^2\right) c_\rho s_\rho + \widehat{\zeta}_{23} c_\nu^2 s^2 c_\sigma s_\sigma\right] \; , \\ \dot{\sigma} &= \frac{C_\kappa y_\tau^2}{16\pi^2}\left[\widehat{\zeta}_{12} s_\nu^2 s^2 c_{(\sigma-\rho)} s_{(\sigma-\rho)} + \widehat{\zeta}_{13} s_\nu^2 s^2 c_\rho s_\rho + \widehat{\zeta}_{23}\left(c_\nu^2 s^2 - c^2\right) c_\sigma s_\sigma\right] \; , \\ \dot{\phi} &= \frac{C_\kappa y_\tau^2}{16\pi^2}\left[\left(c_l^2 - s_l^2\right) c_l^{-1} s_l^{-1} c_\nu s_\nu c \left(\zeta_{13}^{-1} c_\rho s_{(\rho-\phi)} - \zeta_{13} s_\rho c_{(\rho-\phi)} - \zeta_{23}^{-1} c_\sigma s_{(\sigma-\phi)} + \zeta_{23} s_\sigma c_{(\sigma-\phi)}\right)\right. \\ &\quad \left. + \widehat{\zeta}_{12} s^2 c_{(\sigma-\rho)} s_{(\sigma-\rho)} + \widehat{\zeta}_{13}\left(s_\nu^2 - c_\nu^2 c^2\right) c_\rho s_\rho + \widehat{\zeta}_{23}\left(c_\nu^2 - s_\nu^2 c^2\right) c_\sigma s_\sigma\right] \; , \end{aligned} \tag{139}$$

where $\widehat{\zeta}_{ij} \equiv \zeta_{ij}^{-1} - \zeta_{ij} = 4\kappa_i\kappa_j / \left(\kappa_i^2 - \kappa_j^2\right)$, $c_a \equiv \cos a$ and $s_a \equiv \sin a$ (for $a = \rho$, σ, $\sigma - \rho$, $\rho - \phi$ or $\sigma - \phi$).

Some discussions on the basic features of RGEs of three neutrino masses, three flavor mixing angles and three CP-violating phases are in order.

(a) The running behaviors of three neutrino masses m_i (or equivalently κ_i) are essentially identical and determined by α_κ, unless $\tan\beta$ is large enough in the MSSM to make the y_τ^2-associated term competitive with the α_κ term. In our phase convention, $\dot{\kappa}_i$ or $\dot{m}_i$ (for $i = 1, 2, 3$) are independent of the CP-violating phase ϕ.

(b) Among three neutrino mixing angles, only the derivative of θ_ν contains a term proportional to ζ_{12}^{-1}. Note that $\zeta_{ij}^{-1} = (m_i + m_j)^2/\Delta m_{ij}^2$ with $\Delta m_{ij}^2 \equiv m_i^2 - m_j^2$ holds. Current solar and atmospheric neutrino oscillation data yield $\Delta m_{21}^2 \approx 7.7 \times 10^{-5}$ eV^2 and $\left|\Delta m_{32}^2\right| \approx \left|\Delta m_{31}^2\right| \approx 2.4 \times 10^{-3}$ eV^2. So θ_ν is in general more sensitive to radiative corrections than θ_l and θ. The evolution of θ_ν can be suppressed through the fine-tuning of $(\sigma - \rho)$. The smallest neutrino mixing angle θ_l may get radiative

corrections even if its initial value is zero, and thus it can be radiatively generated from other neutrino mixing angles and CP-violating phases.

(c) The running behavior of ϕ is quite different from those of ρ and σ, because it includes a peculiar term proportional to s_l^{-1}. This term, which dominates $\dot{\phi}$ when θ_l is sufficiently small, becomes divergent in the limit $\theta_l \to 0$. Indeed, ϕ is not well-defined if θ_l is exactly vanishing. But both θ_l and ϕ can be radiatively generated. We may require that $\dot{\phi}$ remain finite when θ_l approaches zero, implying that the following necessary condition can be extracted from the expression of $\dot{\phi}$ in Eq. (139):

$$\zeta_{13}^{-1} c_\rho s_{(\rho-\phi)} - \zeta_{13} s_\rho c_{(\rho-\phi)} - \zeta_{23}^{-1} c_\sigma s_{(\sigma-\phi)} + \zeta_{23} s_\sigma c_{(\sigma-\phi)} = 0 \; . \tag{140}$$

Note that the initial value of θ_l, if it is exactly zero or extremely small, may immediately drive ϕ to its *quasi-fixed point*. In this case Eq. (140) can be used to understand the relationship between ϕ and two Majorana phases ρ and σ at the quasi-fixed point.

(d) The running behaviors of ρ and σ are relatively mild in comparison with that of ϕ. A remarkable feature of $\dot{\rho}$ and $\dot{\sigma}$ is that they will vanish, if both ρ and σ are initially vanishing. This observation indicates that ρ and σ cannot simultaneously be generated from ϕ via the RGEs.

6 How to Generate Neutrino Masses?

Neutrinos are assumed or required to be massless in the SM, just because the structure of the SM itself is too simple to accommodate massive neutrinos.

- Two fundamentals of the SM are the $SU(2)_{\rm L} \times U(1)_{\rm Y}$ gauge symmetry and the Lorentz invariance. Both of them are mandatory to guarantee that the SM is a consistent quantum field theory.
- The particle content of the SM is rather economical. There are no right-handed neutrinos in the SM, so a Dirac neutrino mass term is not allowed. There is only one Higgs doublet, so a gauge-invariant Majorana mass term is forbidden.
- The SM is a renormalizable quantum field theory. Hence an effective dimension-5 operator, which may give each neutrino a Majorana mass, is absent.

In other words, the SM accidently possesses the $(B-L)$ symmetry which assures three known neutrinos to be exactly massless.

But today's experiments have convincingly indicated the existence of neutrino oscillations. This quantum phenomenon can appear if and only if neutrinos are massive and lepton flavors are mixed, and thus it is a kind of new physics beyond the SM. To generate non-zero but tiny neutrino masses, one or more of the above-mentioned constraints on the SM must be abandoned or relaxed. It is intolerable to abandon the gauge symmetry and Lorentz invariance; otherwise, one would be led astray. Given the framework of the SM as a consistent field theory, its particle content can be modified and (or) its renormalizability can be abandoned to accommodate massive neutrinos. There are several ways to this goal.

6.1 Relaxing the renormalizability

In 1979, Weinberg extended the SM by introducing some higher-dimension operators in terms of the fields of the SM itself:

$$\mathcal{L}_{\rm eff} = \mathcal{L}_{\rm SM} + \frac{\mathcal{L}_{\rm d=5}}{\Lambda} + \frac{\mathcal{L}_{\rm d=6}}{\Lambda^2} + \cdots \; , \tag{141}$$

where Λ denotes the cut-off scale of this effective theory. Within such a framework, the lowest-dimension operator that violates the lepton number (L) is the unique dimension-5 operator $HHLL/\Lambda$. After spontaneous gauge symmetry breaking, this Weinberg operator yields $m_i \sim \langle H \rangle^2/\Lambda$ for neutrino masses, which can be sufficiently small (≤ 1 eV) if Λ is not far away from the scale of grand unified theories ($\Lambda \sim 10^{13}$ GeV for $\langle H \rangle \sim 10^2$ GeV). In this sense we argue that neutrino masses can serve as a low-energy window onto new physics at superhigh energies.

6.2 A pure Dirac neutrino mass term?

Given three right-handed neutrinos, the gauge-invariant and lepton-number-conserving mass terms of charged leptons and neutrinos are

$$-\mathcal{L}_{\rm lepton} = \overline{\ell_{\rm L}} Y_l H E_{\rm R} + \overline{\ell_{\rm L}} Y_\nu \tilde{H} N_{\rm R} + {\rm h.c.} \; , \tag{142}$$

where $\tilde{H} \equiv i\sigma_2 H^*$ is defined and $\ell_{\rm L}$ denotes the left-handed lepton doublet. After spontaneous gauge symmetry breaking, we arrive at the charged-lepton mass matrix $M_l = Y_l v/\sqrt{2}$ and the Dirac neutrino mass matrix $M_\nu = Y_\nu v/\sqrt{2}$ with $v \simeq 246$ GeV. In this case, the smallness of three neutrino masses m_i (for $i = 1, 2, 3$) is attributed to the smallness of three eigenvalues of Y_ν (denoted as y_i for $i = 1, 2, 3$). Then we encounter a transparent hierarchy problem: $y_i/y_e = m_i/m_e \leq 0.5\,{\rm eV}/0.5\,{\rm MeV} \sim 10^{-6}$. Why is y_i so small? There is no explanation at all in this Dirac-mass picture.

A speculative way out is to invoke extra dimensions; namely, the smallness of Dirac neutrino masses is ascribed to the assumption that three right-handed neutrinos have access to one or more extra spatial dimensions. The idea is simply to confine the SM particles onto a brane and to allow $N_{\rm R}$ to travel in the bulk. For example, the wave-function of $N_{\rm R}$ spreads out over the extra dimension y, giving rise to a suppressed Yukawa interaction at $y = 0$ (i.e., the location of the brane):

$$\left[\overline{\ell_{\rm L}} Y_\nu \tilde{H} N_{\rm R}\right]_{y=0} \sim \frac{1}{\sqrt{L}} \left[\overline{\ell_{\rm L}} Y_\nu \tilde{H} N_{\rm R}\right]_{y=L} \; . \tag{143}$$

The magnitude of $1/\sqrt{L}$ is measured by $\Lambda/\Lambda_{\rm Planck}$, and thus it can naturally be small for an effective theory far below the Planck scale.

6.3 Seesaw mechanisms

This approach works at the tree level and reflects the essential spirit of seesaw mechanisms — tiny masses of three known neutrinos are attributed to the existence of heavy degrees of freedom and lepton number violation.

– Type-I seesaw — three heavy right-handed neutrinos are added into the SM and the lepton number is violated by their Majorana mass term:

$$-\mathcal{L}_{\rm lepton} = \overline{\ell_{\rm L}} Y_l H E_{\rm R} + \overline{\ell_{\rm L}} Y_\nu \tilde{H} N_{\rm R} + \frac{1}{2} \overline{N_{\rm R}^{\rm c}} M_{\rm R} N_{\rm R} + {\rm h.c.} \; , \tag{144}$$

where $M_{\rm R}$ is the Majorana mass matrix.

– Type-II seesaw — one heavy Higgs triplet is added into the SM and the lepton number is violated by its interactions with both the lepton doublet and the Higgs doublet:

$$-\mathcal{L}_{\rm lepton} = \overline{\ell_{\rm L}} Y_l H E_{\rm R} + \frac{1}{2} \overline{\ell_{\rm L}} Y_\Delta \Delta i\sigma_2 \ell_{\rm L}^{\rm c} - \lambda_\Delta M_\Delta H^T i\sigma_2 \Delta H + {\rm h.c.} \; , \tag{145}$$

where

$$\Delta \equiv \begin{pmatrix} \Delta^- & -\sqrt{2}\,\Delta^0 \\ \sqrt{2}\,\Delta^{--} & -\Delta^- \end{pmatrix} \tag{146}$$

denotes the $SU(2)_{\rm L}$ Higgs triplet.

– Type-III seesaw — three heavy triplet fermions are added into the SM and the lepton number is violated by their Majorana mass term:

$$-\mathcal{L}_{\rm lepton} = \overline{\ell_{\rm L}} Y_l H E_{\rm R} + \overline{\ell_{\rm L}} \sqrt{2} Y_\Sigma \Sigma^{\rm c} \tilde{H} + \frac{1}{2} {\rm Tr}\left(\overline{\Sigma} M_\Sigma \Sigma^{\rm c}\right) + {\rm h.c.} \; , \tag{147}$$

where

$$\Sigma = \begin{pmatrix} \Sigma^0/\sqrt{2} & \Sigma^+ \\ \Sigma^- & -\Sigma^0/\sqrt{2} \end{pmatrix} \tag{148}$$

denotes the $SU(2)_{\rm L}$ fermion triplet.

Of course, there are a number of variations or combinations of these three typical seesaw mechanisms in the literature.

For each of the above seesaw pictures, one may arrive at the unique dimension-5 Weinberg operator of neutrino masses after integrating out the corresponding heavy degrees of freedom:

$$\frac{\mathcal{L}_{\rm d=5}}{\Lambda} = \begin{cases} \frac{1}{2}\left(Y_\nu M_{\rm R}^{-1} Y_\nu^T\right)_{\alpha\beta} \overline{\ell_{\alpha{\rm L}}} \tilde{H} \tilde{H}^T \ell^{\rm c}_{\beta{\rm L}} + {\rm h.c.} \\ -\frac{\lambda_\Delta}{M_\Delta} \left(Y_\Delta\right)_{\alpha\beta} \overline{\ell_{\alpha{\rm L}}} \tilde{H} \tilde{H}^T \ell^{\rm c}_{\beta{\rm L}} + {\rm h.c.} \\ \frac{1}{2}\left(Y_\Sigma M_\Sigma^{-1} Y_\Sigma^T\right)_{\alpha\beta} \overline{\ell_{\alpha{\rm L}}} \tilde{H} \tilde{H}^T \ell^{\rm c}_{\beta{\rm L}} + {\rm h.c.} \end{cases}$$

corresponding to type-I, type-II and type-III seesaws. After spontaneous gauge symmetry breaking, $\tilde{H}$ achieves its vacuum expectation value $\langle \tilde{H} \rangle = v/\sqrt{2}$ with $v \simeq 246$ GeV. Then we are left with the effective Majorana neutrino mass term for three known neutrinos,

$$-\mathcal{L}_{\rm mass} = \frac{1}{2} \overline{\nu_{\rm L}} M_\nu \nu^{\rm c}_{\rm L} + {\rm h.c.} \; , \tag{149}$$

where the Majorana mass matrix M_ν is given by

$$M_\nu = \begin{cases} -\frac{1}{2} Y_\nu \frac{v^2}{M_{\rm R}} Y_\nu^T & ({\rm Type\ I}) \; , \\ \lambda_\Delta Y_\Delta \frac{v^2}{M_\Delta} & ({\rm Type\ II}) \; , \\ -\frac{1}{2} Y_\Sigma \frac{v^2}{M_\Sigma} Y_\Sigma^T & ({\rm Type\ III}) \; . \end{cases} \tag{150}$$

It becomes obvious that the smallness of M_ν can be attributed to the largeness of $M_{\rm R}$, M_Δ or M_Σ in the seesaw mechanism.

6.4 Radiative origin of neutrino masses

In a seminal paper published in 1972, Weinberg pointed out that "in theories with spontaneously broken gauge symmetries, various masses or mass differences may vanish in zeroth order as a consequence of the representation content of the fields appearing in the Lagrangian. These masses or mass differences can then be calculated as finite higher-order effects." Such a mechanism may allow us to slightly go beyond the SM and radiatively generate tiny neutrino masses. A typical example is the well-known Zee model,

$$-\mathcal{L}_{\rm lepton} = \overline{\ell_{\rm L}} Y_l H E_{\rm R} + \overline{\ell_{\rm L}} Y_S S^- i\sigma_2 l^{\rm c}_{\rm L} + \tilde{\Phi}^T F S^+ i\sigma_2 \tilde{H} + {\rm h.c.} \; , \tag{151}$$

where $S^\pm$ are charged $SU(2)_{\rm L}$ singlet scalars, Φ denotes a new $SU(2)_{\rm L}$ doublet scalar which has the same quantum number as the SM Higgs doublet H, Y_S is an anti-symmetric matrix, and F represents a mass. Without loss of generality, we choose the basis of $M_l = Y_l \langle H \rangle = {\rm Diag}\{m_e, m_\mu, m_\tau\}$. In this model neutrinos are massless at the tree level, but their masses can radiatively be generated via the one-loop corrections. Given $M_S \gg M_H \sim M_\Phi \sim F$ and $\langle \Phi \rangle \sim \langle H \rangle$, the elements of the effective mass matrix of three light Majorana neutrinos are

$$\left(M_\nu\right)_{\alpha\beta} \; \sim \; \frac{M_H}{16\pi^2} \cdot \frac{m_\alpha^2 - m_\beta^2}{M_S^2} \left(Y_S\right)_{\alpha\beta} \; , \tag{152}$$

where α and β run over e, μ and τ. The smallness of M_ν is therefore ascribed to the smallness of Y_S and $(m_\alpha^2 - m_\beta^2)/M_S^2$. Although the original version of the Zee model is disfavored by current experimental data on neutrino oscillations, its extensions or variations at the one-loop or two-loop level can survive.

7 On the Scales of Seesaw Mechanisms

As we have seen, the key point of a seesaw mechanism is to ascribe the smallness of neutrino masses to the existence of some new degrees of freedom heavier than the Fermi scale $v \simeq 246$ GeV, such as heavy Majorana neutrinos or heavy Higgs bosons. The energy scale where a seesaw mechanism works is crucial, because it is relevant to whether this mechanism is theoretically natural and experimentally testable. Between Fermi and Planck scales, there might exist two other fundamental scales: one is the scale of a grand unified theory (GUT) at which strong, weak and electromagnetic forces can be unified, and the other is the TeV scale at which the unnatural gauge hierarchy problem of the SM can be solved or at least softened by a kind of new physics.

7.1 How about a very low seesaw scale?

In reality, however, there is no direct evidence for a high or extremely high seesaw scale. Hence eV-, keV-, MeV- and GeV-scale seesaws are all possible, at least in principle, and they are technically natural in the sense that their lepton-number-violating mass terms are naturally small according to 't Hooft's naturalness criterion — "At any energy scale μ, a set of parameters $\alpha_i(\mu)$ describing a system can be small, if and only if, in the limit $\alpha_i(\mu) \to 0$ for each of these parameters, the system exhibits an enhanced symmetry." But there are several potential problems associated with low-scale seesaws: (a) a low-scale seesaw does not give any obvious connection to a theoretically well-justified fundamental physical scale (such as the Fermi scale, the TeV scale, the GUT scale or the Planck scale); (b) the neutrino Yukawa couplings in a low-scale seesaw model turn out to be tiny, giving no actual explanation of why the masses of three known neutrinos are so small; and (c) in general, a very low seesaw scale does not allow the "canonical" thermal leptogenesis mechanism to work.

7.2 Seesaw-induced hierarchy problem

Many theorists argue that the conventional seesaw scenarios are natural because their scales (i.e., the masses of heavy degrees of freedom) are close to the GUT scale. This argument is reasonable on the one hand, but it reflects the drawbacks of the conventional seesaw models on the other hand. In other words, the conventional seesaw models have no direct experimental testability and involve a potential hierarchy problem. The latter is usually spoke of when two largely different energy scales exist in a model, but there is no symmetry to stabilize the low-scale physics suffering from large corrections coming from the high-scale physics.

Such a seesaw-induced fine-tuning problem means that the SM Higgs mass is very sensitive to quantum corrections from the heavy degrees of freedom in a seesaw mechanism. For example,

$$\delta M_H^2 = \begin{cases} -\dfrac{y_i^2}{8\pi^2}\left(\Lambda^2 + M_i^2 \ln \dfrac{M_i^2}{\Lambda^2}\right) & \text{(I)} \\ \dfrac{3}{16\pi^2}\left[\lambda_3\left(\Lambda^2 + M_\Delta^2 \ln \dfrac{M_\Delta^2}{\Lambda^2}\right) + 4\lambda_\Delta^2 M_\Delta^2 \ln \dfrac{M_\Delta^2}{\Lambda^2}\right] & \text{(II)} \\ -\dfrac{3y_i^2}{8\pi^2}\left(\Lambda^2 + M_i^2 \ln \dfrac{M_i^2}{\Lambda^2}\right) & \text{(III)} \end{cases}$$

in three typical seesaw scenarios, where Λ is the regulator cut-off, y_i and M_i (for $i = 1, 2, 3$) stand respectively for the eigenvalues of Y_ν (or Y_Σ) and $M_{\rm R}$ (or M_Σ), and the contributions proportional to v^2 and M_H^2 have been omitted. The above results show a quadratic sensitivity to the new scale which is characteristic of the seesaw model, implying that a high degree of fine-tuning would be necessary to accommodate the experimental data on M_H if the seesaw scale is much larger than v (or the Yukawa couplings are not extremely fine-tuned in type-I and type-III seesaws). Taking the type-I seesaw scenario for illustration, we assume $\Lambda \sim M_i$ and require $|\delta M_H^2| \leq 0.1$ TeV^2. Then the above equation leads us

to the following rough estimate:

$$M_i \sim \left[\frac{(2\pi v)^2 |\delta M_H^2|}{m_i}\right]^{1/3} \leq 10^7 \text{GeV} \left[\frac{0.2 \text{ eV}}{m_i}\right]^{1/3} \left[\frac{|\delta M_H^2|}{0.1 \text{ TeV}^2}\right]^{1/3} . \tag{153}$$

This naive result indicates that a hierarchy problem will arise if the masses of heavy Majorana neutrinos are larger than about 10^7 GeV in the type-I seesaw scheme. Because of $m_i \sim y_i^2 v^2/(2M_i)$, the bound $M_i \leq 10^7$ GeV implies $y_i \sim \sqrt{2m_i M_i}/v \leq 2.6 \times 10^{-4}$ for $m_i \sim 0.2$ eV. Such a small magnitude of y_i seems to be a bit unnatural in the sense that the conventional seesaw idea attributes the smallness of m_i to the largeness of M_i other than the smallness of y_i.

There are two possible ways out of this impasse: one is to appeal for the supersymmetry, and the other is to lower the seesaw scale. We shall follow the second way to discuss the TeV seesaw mechanisms which do not suffer from the above-mentioned hierarchy problem.

7.3 Why are the TeV seesaws interesting?

There are several reasons for people to expect some new physics at the TeV scale. This kind of new physics should be able to stabilize the Higgs-boson mass and hence the electroweak scale; in other words, it should be able to solve or soften the unnatural gauge hierarchy problem. It has also been argued that the weakly-interacting particle candidates for dark matter should weigh about one TeV or less. If the TeV scale is really a fundamental scale, may we argue that the TeV seesaws are natural? Indeed, we are reasonably motivated to speculate that possible new physics existing at the TeV scale and responsible for the electroweak symmetry breaking might also be responsible for the origin of neutrino masses. It is interesting and meaningful in this sense to investigate and balance the "naturalness" and "testability" of TeV seesaws at the energy frontier set by the LHC.

As a big bonus of the conventional (type-I) seesaw mechanism, the thermal leptogenesis mechanism provides us with an elegant dynamic picture to interpret the cosmological matter-antimatter asymmetry characterized by the observed ratio of baryon number density to photon number density, $\eta_B \equiv n_B/n_\gamma = (6.1 \pm 0.2) \times 10^{10}$. When heavy Majorana neutrino masses are down to the TeV scale, the Yukawa couplings should be reduced by more than six orders of magnitude so as to generate tiny masses for three known neutrinos via the type-I seesaw and satisfy the out-of-equilibrium condition, but the CP-violating asymmetries of heavy Majorana neutrino decays can still be enhanced by the resonant effects in order to account for η_B. This "resonant leptogenesis" scenario might work in a specific TeV seesaw model.

Is there a TeV Noah's Ark which can naturally and simultaneously accommodate the seesaw idea, the leptogenesis picture and the collider signatures? We are most likely not so lucky and should not be too ambitious at present. In the following we shall concentrate on the TeV seesaws themselves and their possible collider signatures and low-energy consequences.

8 TeV Seesaws: Natural and Testable?

The neutrino mass terms in three typical seesaw mechanisms have been given before. Without loss of generality, we choose the basis in which the mass eigenstates of three charged leptons are identified with their flavor eigenstates.

8.1 Type-I seesaw

Given $M_{\rm D} = Y_\nu v/\sqrt{2}$, the approximate type-I seesaw formula in Eq. (150) can be rewritten as $M_\nu = -M_{\rm D} M_{\rm R}^{-1} M_{\rm D}^T$. Note that the 3×3 light neutrino mixing matrix V is not exactly unitary in this seesaw scheme, and its deviation from unitarity is of $\mathcal{O}(M_{\rm D}^2/M_{\rm R}^2)$. Let us consider two interesting possibilities. (1) $M_{\rm D} \sim \mathcal{O}(10^2)$ GeV and $M_{\rm R} \sim \mathcal{O}(10^{15})$ GeV to get $M_\nu \sim \mathcal{O}(10^{-2})$ eV. In this conventional and

natural case, $M_{\rm D}/M_{\rm R} \sim \mathcal{O}(10^{-13})$ holds. Hence the non-unitarity of V is only at the $\mathcal{O}(10^{-26})$ level, too small to be observed. (2) $M_{\rm D} \sim \mathcal{O}(10^2)$ GeV and $M_{\rm R} \sim \mathcal{O}(10^3)$ GeV to get $M_\nu \sim \mathcal{O}(10^{-2})$ eV. In this *unnatural* case, a significant "structural cancellation" has to be imposed on the textures of $M_{\rm D}$ and $M_{\rm R}$. Because of $M_{\rm D}/M_{\rm R} \sim \mathcal{O}(0.1)$, the non-unitarity of V can reach the percent level and may lead to observable effects.

Now we discuss how to realize the above "structural cancellation" for the type-I seesaw mechanism at the TeV scale. For the sake of simplicity, we take the basis of $M_{\rm R} = {\rm Diag}\{M_1, M_2, M_3\}$ for three heavy Majorana neutrinos (N_1, N_2, N_3). It is well known that M_ν vanishes if

$$M_{\rm D} = m \begin{pmatrix} y_1 & y_2 & y_3 \\ \alpha y_1 & \alpha y_2 & \alpha y_3 \\ \beta y_1 & \beta y_2 & \beta y_3 \end{pmatrix} \; , \quad \sum_{i=1}^{3} \frac{y_i^2}{M_i} = 0 \tag{154}$$

simultaneously hold. Tiny neutrino masses can be generated from tiny corrections to the texture of $M_{\rm D}$ in Eq. (154). For example, $M'_{\rm D} = M_{\rm D} - \epsilon X_{\rm D}$ with $M_{\rm D}$ given above and ϵ being a small dimensionless parameter (i.e., $|\epsilon| \ll 1$) yields

$$M'_\nu = -M'_{\rm D} M_{\rm R}^{-1} M'^T_{\rm D} \simeq \epsilon \left(M_{\rm D} M_{\rm R}^{-1} X_{\rm D}^T + X_{\rm D} M_{\rm R}^{-1} M_{\rm D}^T \right) \; , \tag{155}$$

from which $M'_\nu \sim \mathcal{O}(10^{-2})$ eV can be obtained by adjusting the size of ϵ.

A lot of attention has recently been paid to a viable type-I seesaw model and its collider signatures at the TeV scale. At least the following lessons can be learnt:

- Two necessary conditions must be satisfied in order to test a type-I seesaw model at the LHC: (a) M_i are of $\mathcal{O}(1)$ TeV or smaller; and (b) the strength of light-heavy neutrino mixing (i.e., $M_{\rm D}/M_{\rm R}$) is large enough. Otherwise, it would be impossible to produce and detect N_i at the LHC.
- The collider signatures of N_i are essentially decoupled from the mass and mixing parameters of three light neutrinos ν_i. For instance, the small parameter ϵ in Eq. (155) has nothing to do with the ratio $M_{\rm D}/M_{\rm R}$.
- The non-unitarity of V might lead to some observable effects in neutrino oscillations and other lepton-flavor-violating or lepton-number-violating processes, if $M_{\rm D}/M_{\rm R} \leq \mathcal{O}(0.1)$ holds.
- The clean LHC signatures of heavy Majorana neutrinos are the $\Delta L = 2$ like-sign dilepton events, such as $pp \to W^{*\pm} W^{*\pm} \to \mu^\pm \mu^\pm jj$ and $pp \to W^{*\pm} \to \mu^\pm N_i \to \mu^\pm \mu^\pm jj$ (a dominant channel due to the resonant production of N_i).

Some instructive and comprehensive analyses of possible LHC events for a single heavy Majorana neutrino have recently been done, but they only serve for illustration because such a simplified type-I seesaw scenario is actually unrealistic.

8.2 Type-II seesaw

The type-II seesaw formula $M_\nu = Y_\Delta v_\Delta = \lambda_\Delta Y_\Delta v^2/M_\Delta$ has been given in Eq. (150). Note that the last term of Eq. (145) violates both L and $B - L$, and thus the smallness of λ_Δ is naturally allowed according to 't Hooft's naturalness criterion (i.e., setting $\lambda_\Delta = 0$ will increase the symmetry of $\mathcal{L}_{\rm lepton}$). Given $M_\Delta \sim \mathcal{O}(1)$ TeV, for example, this seesaw mechanism works to generate $M_\nu \sim \mathcal{O}(10^{-2})$ eV provided $\lambda_\Delta Y_\Delta \sim \mathcal{O}(10^{-12})$ holds. The neutrino mixing matrix V is exactly unitary in the type-II seesaw mechanism, simply because the heavy degrees of freedom do not mix with the light ones.

There are totally seven physical Higgs bosons in the type-II seesaw scheme: doubly-charged H^{++} and H^{--}, singly-charged H^+ and H^-, neutral A^0 (CP-odd), and neutral h^0 and H^0 (CP-even), where h^0 is the SM-like Higgs boson. Except for $M^2_{h^0}$, we get a quasi-degenerate mass spectrum for other scalars: $M^2_{H^{\pm\pm}} = M^2_\Delta \approx M^2_{H^0} \approx M^2_{H^\pm} \approx M^2_{A^0}$. As a consequence, the decay channels $H^{\pm\pm} \to W^\pm H^\pm$ and

$H^{\pm\pm} \to H^\pm H^\pm$ are kinematically forbidden. The production of $H^{\pm\pm}$ at the LHC is mainly through $q\bar{q} \to \gamma^*, Z^* \to H^{++}H^{--}$ and $q\bar{q}' \to W^* \to H^{\pm\pm}H^\mp$ processes, which do not rely on the small Yukawa couplings.

The typical collider signatures in this seesaw scenario are the lepton-number-violating $H^{\pm\pm} \to l^\pm_\alpha l^\pm_\beta$ decays as well as $H^+ \to l^+_\alpha \overline{\nu}$ and $H^- \to l^-_\alpha \nu$ decays. Their branching ratios

$$\begin{aligned} \mathcal{B}(H^{\pm\pm} \to l^\pm_\alpha l^\pm_\beta) &= \frac{|(M_\nu)_{\alpha\beta}|^2 \left(2 - \delta_{\alpha\beta}\right)}{\sum\limits_{\rho,\sigma} |(M_\nu)_{\rho\sigma}|^2} \; , \\ \mathcal{B}(H^+ \to l^+_\alpha \overline{\nu}) &= \frac{\sum\limits_\beta |(M_\nu)_{\alpha\beta}|^2}{\sum\limits_{\rho,\sigma} |(M_\nu)_{\rho\sigma}|^2} \end{aligned} \tag{156}$$

are closely related to the masses, flavor mixing angles and CP-violating phases of three light neutrinos, because $M_\nu = V \widehat{M}_\nu V^T$ with $\widehat{M}_\nu = {\rm Diag}\{m_1, m_2, m_3\}$ holds. Some detailed analyses of such decay modes together with the LHC signatures of $H^{\pm\pm}$ and $H^\pm$ bosons have been done in the literature.

It is worth pointing out that the following dimension-6 operator can easily be derived from the type-II seesaw mechanism,

$$\frac{\mathcal{L}_{\rm d=6}}{\Lambda^2} = -\frac{(Y_\Delta)_{\alpha\beta} \, (Y_\Delta)^\dagger_{\rho\sigma}}{4 M^2_\Delta} (\overline{\ell_{\alpha {\rm L}}} \gamma^\mu \ell_{\sigma {\rm L}}) (\overline{\ell_{\beta {\rm L}}} \gamma_\mu \ell_{\rho {\rm L}}) \; , \tag{157}$$

which has two immediate low-energy effects: the non-standard interactions of neutrinos and the lepton-flavor-violating interactions of charged leptons. An analysis of such effects provides us with some preliminary information:

- The magnitudes of non-standard interactions of neutrinos and the widths of lepton-flavor-violating tree-level decays of charged leptons are both dependent on neutrino masses m_i and flavor-mixing and CP-violating parameters of V.
- For a long-baseline neutrino oscillation experiment, the neutrino beam encounters the earth matter and the electron-type non-standard interaction contributes to the matter potential.
- At a neutrino factory, the lepton-flavor-violating processes $\mu^- \to e^- \nu_e \overline{\nu}_\mu$ and $\mu^+ \to e^+ \overline{\nu}_e \nu_\mu$ could cause some wrong-sign muons at a near detector.

Current experimental constraints tell us that such low-energy effects are very small, but they might be experimentally accessible in the future precision measurements.

8.3 Type-(I+II) seesaw

The type-(I+II) seesaw mechanism can be achieved by combining the neutrino mass terms in Eqs. (144) and (145). After spontaneous gauge symmetry breaking, we are left with the overall neutrino mass term

$$-\mathcal{L}_{\rm mass} = \frac{1}{2} \overline{\left(\nu_{\rm L} N^{\rm c}_{\rm R}\right)} \begin{pmatrix} M_{\rm L} & M_{\rm D} \\ M^T_{\rm D} & M_{\rm R} \end{pmatrix} \begin{pmatrix} \nu^{\rm c}_{\rm L} \\ N_{\rm R} \end{pmatrix} + {\rm h.c.} \; , \tag{158}$$

where $M_{\rm D} = Y_\nu v/\sqrt{2}$ and $M_{\rm L} = Y_\Delta v_\Delta$ with $\langle H \rangle \equiv v/\sqrt{2}$ and $\langle \Delta \rangle \equiv v_\Delta$ corresponding to the vacuum expectation values of the neutral components of the Higgs doublet H and the Higgs triplet Δ. The 6×6 mass matrix in Eq. (158) is symmetric and can be diagonalized by the unitary transformation done in Eq. (28); i.e.,

$$\begin{pmatrix} V & R \\ S & U \end{pmatrix}^\dagger \begin{pmatrix} M_{\rm L} & M_{\rm D} \\ M^T_{\rm D} & M_{\rm R} \end{pmatrix} \begin{pmatrix} V & R \\ S & U \end{pmatrix}^* = \begin{pmatrix} \widehat{M}_\nu & \mathbf{0} \\ \mathbf{0} & \widehat{M}_N \end{pmatrix} \; , \tag{159}$$

where $\widehat{M}_\nu = {\rm Diag}\{m_1, m_2, m_3\}$ and $\widehat{M}_N = {\rm Diag}\{M_1, M_2, M_3\}$. Needless to say, $V^\dagger V + S^\dagger S = V V^\dagger + R R^\dagger = {\bf 1}$ holds as a consequence of the unitarity of this transformation. Hence V, the flavor mixing matrix of light Majorana neutrinos, must be non-unitary if R and S are non-zero.

In the leading-order approximation, the type-(I+II) seesaw formula reads as

$$M_\nu \approx M_{\rm L} - M_{\rm D} M_{\rm R}^{-1} M_{\rm D}^T \; . \tag{160}$$

Hence type-I and type-II seesaws can be regarded as two extreme cases of the type-(I+II) seesaw. Note that two mass terms in Eq. (160) are possibly comparable in magnitude. If both of them are small, their contributions to M_ν may have significant interference effects which make it practically impossible to distinguish between type-II and type-(I+II) seesaws; but if both of them are large, their contributions to M_ν must be destructive. The latter case unnaturally requires a significant cancellation between two big quantities in order to obtain a small quantity, but it is interesting in the sense that it may give rise to possibly observable collider signatures of heavy Majorana neutrinos.

Let me briefly describe a particular type-(I+II) seesaw model and comment on its possible LHC signatures. First, we assume that both M_i and M_Δ are of $\mathcal{O}(1)$ TeV. Then the production of $H^{\pm\pm}$ and $H^\pm$ bosons at the LHC is guaranteed, and their lepton-number-violating signatures will probe the Higgs triplet sector of the type-(I+II) seesaw mechanism. On the other hand, $\mathcal{O}(M_{\rm D}/M_{\rm R}) \leq \mathcal{O}(0.1)$ is possible as a result of $\mathcal{O}(M_{\rm R}) \sim \mathcal{O}(1)$ TeV and $\mathcal{O}(M_{\rm D}) \leq \mathcal{O}(v)$, such that appreciable signatures of N_i can be achieved at the LHC. Second, the small mass scale of M_ν implies that the relation $\mathcal{O}(M_{\rm L}) \sim \mathcal{O}(M_{\rm D} M_{\rm R}^{-1} M_{\rm D}^T)$ must hold. In other words, it is the significant but incomplete cancellation between $M_{\rm L}$ and $M_{\rm D} M_{\rm R}^{-1} M_{\rm D}^T$ terms that results in the non-vanishing but tiny masses for three light neutrinos. We admit that dangerous radiative corrections to two mass terms of M_ν require a delicate fine-tuning of the cancellation at the loop level. But this scenario allows us to reconstruct $M_{\rm L}$ via the excellent approximation $M_{\rm L} = V \widehat{M}_\nu V^T + R \widehat{M}_N R^T \approx R \widehat{M}_N R^T$, such that the elements of the Yukawa coupling matrix Y_Δ read as follows:

$$(Y_\Delta)_{\alpha\beta} \; = \; \frac{(M_{\rm L})_{\alpha\beta}}{v_\Delta} \approx \sum_{i=1}^{3} \frac{R_{\alpha i} R_{\beta i} M_i}{v_\Delta} \; , \tag{161}$$

where the subscripts α and β run over e, μ and τ. This result implies that the leptonic decays of $H^{\pm\pm}$ and $H^\pm$ bosons depend on both R and M_i, which actually determine the production and decays of N_i. Thus we have established an interesting correlation between the singly- or doubly-charged Higgs bosons and the heavy Majorana neutrinos. To observe the correlative signatures of $H^\pm$, $H^{\pm\pm}$ and N_i at the LHC will serve for a direct test of this type-(I+II) seesaw model.

8.4 Type-III seesaw

The lepton mass terms in the type-III seesaw scheme have already been given in Eq. (147). After spontaneous gauge symmetry breaking, we are left with

$$\begin{aligned} -\mathcal{L}_{\rm mass} &= \frac{1}{2} \overline{\left(\nu_{\rm L} \; \Sigma^0\right)} \begin{pmatrix} {\bf 0} & M_{\rm D} \\ M_{\rm D}^T & M_\Sigma \end{pmatrix} \begin{pmatrix} \nu_{\rm L}^{\rm c} \\ \Sigma^{0{\rm c}} \end{pmatrix} + {\rm h.c.} \; , \\ -\mathcal{L}'_{\rm mass} &= \overline{\left(e_{\rm L} \; \Psi_{\rm L}\right)} \begin{pmatrix} M_l & \sqrt{2} M_{\rm D} \\ {\bf 0} & M_\Sigma \end{pmatrix} \begin{pmatrix} E_{\rm R} \\ \Psi_{\rm R} \end{pmatrix} + {\rm h.c.} \; , \end{aligned} \tag{162}$$

respectively, for neutral and charged fermions, where $M_l = Y_l v/\sqrt{2}$, $M_{\rm D} = Y_\Sigma v/\sqrt{2}$, and $\Psi = \Sigma^- + \Sigma^{+{\rm c}}$. The symmetric 6×6 neutrino mass matrix can be diagonalized by the following unitary transformation:

$$\begin{pmatrix} V & R \\ S & U \end{pmatrix}^\dagger \begin{pmatrix} {\bf 0} & M_{\rm D} \\ M_{\rm D}^T & M_\Sigma \end{pmatrix} \begin{pmatrix} V & R \\ S & U \end{pmatrix}^* = \begin{pmatrix} \widehat{M}_\nu & {\bf 0} \\ {\bf 0} & \widehat{M}_\Sigma \end{pmatrix} \; , \tag{163}$$

where $\widehat{M}_\nu = {\rm Diag}\{m_1, m_2, m_3\}$ and $\widehat{M}_\Sigma = {\rm Diag}\{M_1, M_2, M_3\}$. In the leading-order approximation, this diagonalization yields the type-III seesaw formula $M_\nu = -M_{\rm D} M_\Sigma^{-1} M_{\rm D}^T$, which is equivalent to the one derived from the effective dimension-5 operator in Eq. (150). Let us use one sentence to comment on the similarities and differences between type-I and type-III seesaw mechanisms: the non-unitarity of the 3×3 neutrino mixing matrix V has appeared in both cases, although the modified couplings between the Z^0 boson and three light neutrinos differ and the non-unitary flavor mixing is also present in the couplings between the Z^0 boson and three charged leptons in the type-III seesaw scenario.

At the LHC, the typical lepton-number-violating signatures of the type-III seesaw mechanism can be $pp \to \Sigma^+\Sigma^0 \to l_\alpha^+ l_\beta^+ + Z^0 W^-(\to 4j)$ and $pp \to \Sigma^-\Sigma^0 \to l_\alpha^- l_\beta^- + Z^0 W^+(\to 4j)$ processes. A detailed analysis of such collider signatures have been done in the literature. As for the low-energy phenomenology, a consequence of this seesaw scenario is the non-unitarity of the 3×3 flavor mixing matrix N ($\approx V$) in both charged- and neutral-current interactions. Current experimental bounds on the deviation of $NN^\dagger$ from the identity matrix are at the 0.1% level, much stronger than those obtained in the type-I seesaw scheme, just because the flavor-changing processes with charged leptons are allowed at the tree level in the type-III seesaw mechanism.

8.5 Inverse and multiple seesaws

Given the naturalness and testability as two prerequisites, the double or inverse seesaw mechanism is another interesting possibility of generating tiny neutrino masses at the TeV scale. The idea of this seesaw picture is to add three heavy right-handed neutrinos $N_{\rm R}$, three SM gauge-singlet neutrinos $S_{\rm R}$ and one Higgs singlet Φ into the SM, such that the gauge-invariant lepton mass terms can be written as

$$-\mathcal{L}_{\rm lepton} = \overline{l_{\rm L}} Y_l H E_{\rm R} + \overline{l_{\rm L}} Y_\nu \tilde{H} N_{\rm R} + \overline{N_{\rm R}^{\rm c}} Y_S \Phi S_{\rm R} + \frac{1}{2} \overline{S_{\rm R}^{\rm c}} \mu S_{\rm R} + {\rm h.c.} \; , \tag{164}$$

where the μ-term is naturally small according to 't Hooft's naturalness criterion, because it violates the lepton number. After spontaneous gauge symmetry breaking, the overall neutrino mass term turns out to be

$$-\mathcal{L}_{\rm mass} \; = \; \frac{1}{2} \overline{\left(\nu_{\rm L} \; N_{\rm R}^{\rm c} \; S_{\rm R}^{\rm c}\right)} \begin{pmatrix} \mathbf{0} & M_{\rm D} & \mathbf{0} \\ M_{\rm D}^T & \mathbf{0} & M_S \\ \mathbf{0} & M_S^T & \mu \end{pmatrix} \begin{pmatrix} \nu_{\rm L}^{\rm c} \\ N_{\rm R} \\ S_{\rm R} \end{pmatrix} \; , \tag{165}$$

where $M_{\rm D} = Y_\nu \langle H \rangle$ and $M_S = Y_S \langle \Phi \rangle$. A diagonalization of the symmetric 9×9 matrix $\mathcal{M}$ leads us to the effective light neutrino mass matrix

$$M_\nu \approx M_{\rm D} \frac{1}{M_S^T} \mu \frac{1}{M_S} M_{\rm D}^T \tag{166}$$

in the leading-order approximation. Hence the smallness of M_ν can be attributed to both the smallness of μ itself and the doubly-suppressed $M_{\rm D}/M_S$ term for $M_{\rm D} \ll M_S$. For example, $\mu \sim \mathcal{O}(1)$ keV and $M_{\rm D}/M_S \sim \mathcal{O}(10^{-2})$ naturally give rise to a sub-eV M_ν. One has $M_\nu = \mathbf{0}$ in the limit $\mu \to \mathbf{0}$, which reflects the restoration of the slightly-broken lepton number. The heavy sector consists of three pairs of pseudo-Dirac neutrinos whose CP-conjugated Majorana components have a tiny mass splitting characterized by the order of μ.

Going beyond the canonical (type-I) and inverse seesaw mechanisms, one may build the so-called "multiple" seesaw mechanisms to further lower the seesaw scales.

9 Non-unitary Neutrino Mixing

It is worth remarking that the charged-current interactions of light and heavy Majorana neutrinos are not completely independent in either the type-I seesaw or the type-(I+II) seesaw. The standard charged-current interactions of ν_i and N_i are already given in Eq. (34), where V is just the light neutrino mixing

matrix responsible for neutrino oscillations, and R describes the strength of charged-current interactions between (e, μ, τ) and (N_1, N_2, N_3). Since V and R belong to the same unitary transformation done in Eq. (28) or Eq. (159), they must be correlated with each other and their correlation signifies an important relationship between neutrino physics and collider physics.

It can be shown that V and R share nine rotation angles (θ_{i4}, θ_{i5} and θ_{i6} for $i = 1$, 2 and 3) and nine phase angles (δ_{i4}, δ_{i5} and δ_{i6} for $i = 1$, 2 and 3). To see this point clearly, let us decompose V into $V = AV_0$, where V_0 is the standard (unitary) parametrization of the 3×3 PMNS matrix in which three CP-violating phases δ_{ij} (for $ij = 12, 13, 23$) are associated with s_{ij} (i.e., $c_{ij} \equiv \cos\theta_{ij}$ and $\hat{s}_{ij} \equiv e^{i\delta_{ij}} \sin\theta_{ij}$). Because of $VV^\dagger = AA^\dagger = \mathbf{1} - RR^\dagger$, it is obvious that $V \to V_0$ in the limit of $A \to \mathbf{1}$ (or equivalently, $R \to \mathbf{0}$). Considering the fact that the non-unitarity of V must be a small effect (at most at the percent level as constrained by current neutrino oscillation data and precision electroweak data), we expect $s_{ij} \leq \mathcal{O}(0.1)$ (for $i = 1, 2, 3$ and $j = 4, 5, 6$) to hold. Then we obtain

$$R = \begin{pmatrix} \hat{s}^*_{14} & \hat{s}^*_{15} & \hat{s}^*_{16} \\ \hat{s}^*_{24} & \hat{s}^*_{25} & \hat{s}^*_{26} \\ \hat{s}^*_{34} & \hat{s}^*_{35} & \hat{s}^*_{36} \end{pmatrix} \tag{167}$$

as an excellent approximations. A striking consequence of the non-unitarity of V is the loss of universality for the Jarlskog invariants of CP violation, $J^{ij}_{\alpha\beta} \equiv \mathrm{Im}(V_{\alpha i} V_{\beta j} V^*_{\alpha j} V^*_{\beta i})$, where the Greek indices run over (e, μ, τ) and the Latin indices run over $(1, 2, 3)$. For example, the extra CP-violating phases of V are possible to give rise to a significant asymmetry between $\nu_\mu \to \nu_\tau$ and $\overline{\nu}_\mu \to \overline{\nu}_\tau$ oscillations.

The probability of $\nu_\alpha \to \nu_\beta$ oscillations in vacuum, defined as $P_{\alpha\beta}$, is given by

$$P_{\alpha\beta} = \frac{\sum\limits_i |V_{\alpha i}|^2 |V_{\beta i}|^2 + 2 \sum\limits_{i<j} \mathrm{Re}\left(V_{\alpha i} V_{\beta j} V^*_{\alpha j} V^*_{\beta i}\right) \cos\Delta_{ij} - \sum\limits_{i<j} J^{ij}_{\alpha\beta} \sin\Delta_{ij}}{\left(VV^\dagger\right)_{\alpha\alpha} \left(VV^\dagger\right)_{\beta\beta}} \; , \tag{168}$$

where $\Delta_{ij} \equiv \Delta m^2_{ij} L/(2E)$ with $\Delta m^2_{ij} \equiv m^2_i - m^2_j$, E being the neutrino beam energy and L being the baseline length. If V is exactly unitary (i.e., $A = \mathbf{1}$ and $V = V_0$), the denominator of Eq. (168) will become unity and the conventional formula of $P_{\alpha\beta}$ will be reproduced. Note that $\nu_\mu \to \nu_\tau$ and $\overline{\nu}_\mu \to \overline{\nu}_\tau$ oscillations may serve as a good tool to probe possible signatures of non-unitary CP violation. To illustrate this point, we consider a short- or medium-baseline neutrino oscillation experiment with $|\sin\Delta_{13}| \sim |\sin\Delta_{23}| \gg |\sin\Delta_{12}|$, in which the terrestrial matter effects are expected to be insignificant or negligibly small. Then the dominant CP-conserving and CP-violating terms of $P(\nu_\mu \to \nu_\tau)$ and $P(\overline{\nu}_\mu \to \overline{\nu}_\tau)$ are

$$\begin{aligned} P(\nu_\mu \to \nu_\tau) &\approx \sin^2 2\theta_{23} \sin^2 \frac{\Delta_{23}}{2} - 2\left(J^{23}_{\mu\tau} + J^{13}_{\mu\tau}\right) \sin\Delta_{23} \; , \\ P(\overline{\nu}_\mu \to \overline{\nu}_\tau) &\approx \sin^2 2\theta_{23} \sin^2 \frac{\Delta_{23}}{2} + 2\left(J^{23}_{\mu\tau} + J^{13}_{\mu\tau}\right) \sin\Delta_{23} \; , \end{aligned} \tag{169}$$

where the good approximation $\Delta_{13} \approx \Delta_{23}$ has been used in view of the experimental fact $|\Delta m^2_{13}| \approx |\Delta m^2_{23}| \gg |\Delta m^2_{12}|$, and the sub-leading and CP-conserving "zero-distance" effect has been omitted. For simplicity, I take V_0 to be the exactly tri-bimaximal mixing pattern (i.e., $\theta_{12} = \arctan(1/\sqrt{2})$, $\theta_{13} = 0$ and $\theta_{23} = \pi/4$ as well as $\delta_{12} = \delta_{13} = \delta_{23} = 0$) and then arrive at

$$2\left(J^{23}_{\mu\tau} + J^{13}_{\mu\tau}\right) \; \approx \; \sum_{l=4}^{6} s_{2l} s_{3l} \sin\left(\delta_{2l} - \delta_{3l}\right) \; . \tag{170}$$

Given $s_{2l} \sim s_{3l} \sim \mathcal{O}(0.1)$ and $(\delta_{2l} - \delta_{3l}) \sim \mathcal{O}(1)$ (for $l = 4, 5, 6$), this non-trivial CP-violating quantity can reach the percent level. When a long-baseline neutrino oscillation experiment is concerned, however,

the terrestrial matter effects must be taken into account because they might fake the genuine CP-violating signals. As for $\nu_\mu \to \nu_\tau$ and $\overline{\nu}_\mu \to \overline{\nu}_\tau$ oscillations under discussion, the dominant matter effect results from the neutral-current interactions and modifies the CP-violating quantity of Eq. (170) in the following way:

$$2\left(J^{23}_{\mu\tau} + J^{13}_{\mu\tau}\right) \Longrightarrow \sum_{l=4}^{6} s_{2l} s_{3l} \left[\sin\left(\delta_{2l} - \delta_{3l}\right) + A_{\rm NC} L \cos\left(\delta_{2l} - \delta_{3l}\right)\right] \; , \tag{171}$$

where $A_{\rm NC} = G_{\rm F} N_n / \sqrt{2}$ with N_n being the background density of neutrons, and L is the baseline length. It is easy to find $A_{\rm NC} L \sim \mathcal{O}(1)$ for $L \sim 4 \times 10^3$ km.

10 Concluding Remarks

I have briefly described some basic properties of massive neutrinos in an essentially model-independent way in these lectures, which are largely based on the book by Dr. Shun Zhou and myself [1] and on a few review articles or lectures [2]— [6]. It is difficult to cite all the relevant references. I apologize for missing other people's works due to the tight page limit of these proceedings. For the same reason I am unable to write in the cosmological matter-antimatter asymmetry and the leptogenesis mechanism, although they were discussed in my lectures. Here let me just give a few remarks on the naturalness and testability of TeV seesaw mechanisms.

Although the seesaw ideas are elegant, they have to appeal for some or many new degrees of freedom in order to interpret the observed neutrino mass hierarchy and lepton flavor mixing. According to Weinberg's *third law of progress in theoretical physics*, "you may use any degrees of freedom you like to describe a physical system, but if you use the wrong ones, you will be sorry." What could be better?

Anyway, we hope that the LHC might open a new window for us to understand the origin of neutrino masses and the dynamics of lepton number violation. A TeV seesaw might work (*naturalness*?) and its heavy degrees of freedom might show up at the LHC (*testability*?). A bridge between collider physics and neutrino physics is highly anticipated and, if it exists, will lead to rich phenomenology.

I am indebted to the organizers of AEPSHEP 2012 for their invitation and hospitality. This work is supported in part by the National Natural Science Foundation of China under grant No. 11135009.

References

[1] Z.Z. Xing and S. Zhou, *Neutrinos in Particle Physics, Astronomy and Cosmology*, Zhejiang University Press and Springer-Verlag (2011).

[2] Z.Z. Xing, plenary talk given at ICHEP2008; Int. J. Mod. Phys. A **23**, 4255 (2008).

[3] Z.Z. Xing, Prog. Theor. Phys. Suppl. **180**, 112 (2009).

[4] H. Fritzsch and Z.Z. Xing, Prog. Part. Nucl. Phys. **45**, 1 (2000);

[5] Z.Z. Xing, Int. J. Mod. Phys. A **19**, 1 (2004).

[6] Z.Z. Xing, Lectures given at the 2010 Schladming Winter School on *Masses and Constants*, Austria, 2010; published in Nucl. Phys. Proc. Suppl. **203-204**, 82 (2010).

Relativistic Heavy-Ion Collisions

Rajeev S. Bhalerao
Department of Theoretical Physics, Tata Institute of Fundamental Research, Mumbai, India

Abstract
The field of relativistic heavy-ion collisions is introduced to the high-energy physics students with no prior knowledge in this area. The emphasis is on the two most important observables, namely the azimuthal collective flow and jet quenching, and on the role fluid dynamics plays in the interpretation of the data. Other important observables described briefly are constituent quark number scaling, ratios of particle abundances, strangeness enhancement, and sequential melting of heavy quarkonia. Comparison is made of some of the basic heavy-ion results obtained at LHC with those obtained at RHIC. Initial findings at LHC which seem to be in apparent conflict with the accumulated RHIC data are highlighted.

1 Introduction

These are exciting times if one is working in the area of relativistic heavy-ion collisions, with two heavy-ion colliders namely the Relativistic Heavy-Ion Collider (RHIC) at the Brookhaven National Laboratory and the Large Hadron Collider (LHC) at CERN in operation in tandem. Quark-gluon plasma has been discovered at RHIC, but its precise properties are yet to be established. With the phase diagram of strongly interacting matter (QCD phase diagram) also being largely unknown, these are also great times for fresh graduate students to get into this area of research, which is going to remain very active for the next decade at least. The field is maturing as evidenced by the increasing number of text books that are now available [1–9]. Also available are collected review articles; see e.g., [10–12].

This is a fascinating inter-disciplinary area of research at the interface of particle physics and high-energy nuclear physics. It draws heavily from QCD — perturbative, non-perturbative, as well as semiclassical. It has overlaps with thermal field theory, relativistic fluid dynamics, kinetic or transport theory, quantum collision theory, apart from the standard statistical mechanics and thermodynamics. Quark-Gluon Plasma (QGP) at high temperature, T, and vanishing net baryon number density, n_B (or equivalently the corresponding chemical potential, μ_B), is of cosmological interest, while QGP at low T and large n_B is of astrophysical interest. String theorists too have developed interest in this area because of the black hole – fluid dynamics connection.

Students of high-energy physics would know that the science of the 'small' — the elementary particle physics — is deeply intertwined with the science of the 'large' — cosmology — the study of the origin and evolution of the universe. Figure 1 shows the temperature history of the universe starting shortly after the Big Bang. At times $\sim 10\ \mu s$ after the Big Bang, with $T \gtrsim 200$ MeV,[1] the universe was in the state of QGP, and the present-day experiments which collide two relativistic heavy ions — the Little Bang — try to recreate that state of matter in the laboratory for a brief period of time.

Recall the phase diagram (pressure vs temperature) of water, Fig. 2(a). It shows three broad regions separated by phase transition lines, the triple point where all three phases coexist, and the critical point where the vapour pressure curve terminates and two distinct coexisting phases, namely liquid and gas, become identical. All these features are well-established experimentally to a great accuracy. In contrast the QCD phase diagram (Fig. 2(b)) is known only schematically, except for the lattice QCD predictions at vanishing or small μ_B, in particular the prediction of a crossover transition around $T \sim$

[1] In comparison, the temperature and time corresponding to the electroweak transition were $\sim$ 200 GeV and $\sim 10^{-12}$ s, respectively. Note 1 MeV $\simeq 10^{10}$ K.

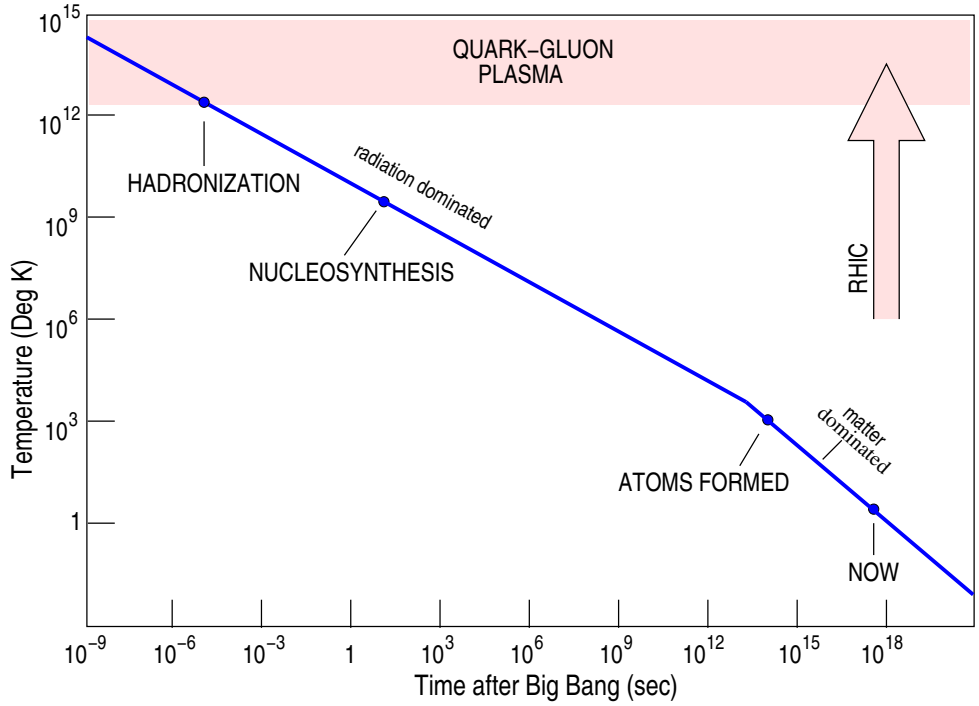


Fig. 1: Temperature history of the universe. The Big Bang and the Little Bang.

150-170 MeV [14, 15] for vanishing μ_B. As arguments based on a variety of models indicate a first-order phase transition as a function of temperature at finite μ_B, one expects the phase transition line to end at a critical point. The existence of the critical point, however, is not established experimentally. Apart from the region of hadrons at the low enough T and μ_B, and the region of quarks and gluons at high T and μ_B, there is also a region characterized by colour superconductivity, at high μ_B and low T [16–18]. However, precise boundaries separating these regions are not known experimentally. Actually the QCD phase diagram may be richer than what is shown in Fig. 2(b) [19]. Before we proceed further, a precise definition of QGP is in order. We follow the definition proposed by the STAR collaboration at RHIC: Quark-Gluon Plasma is defined as a (locally) thermally equilibrated state of matter in which quarks and gluons are deconfined from hadrons, so that they propagate over *nuclear*, rather than merely *nucleonic*, volumes [20]. Note the two essential ingredients of this definition, (a) the constituents of the matter should be quarks and gluons, and (b) the matter should have attained (local)[2] thermal equilibrium. Any claim of discovery of QGP can follow only after these two requirements are shown to be fulfilled unambiguously.

The big idea thus is to map out (quantitatively) the QCD phase diagram [21]. The main theoretical tool at our disposal is, of course, the lattice QCD. Although it allows first-principle calculations, it has technical difficulties for non-vanishing μ_B or n_B. We also have various effective theories and phenomenological models which indeed are the basis of the schematic phase diagram of QCD shown in Fig. 2(b). Experimental tools available to us are the relativistic heavy-ion colliders such as those at BNL and CERN, and the upcoming lower-energy facilities namely Facility for Antiproton and Ion Research (FAIR) at GSI and Nuclotron-based Ion Collider fAcility (NICA) at JINR. Apart from these terrestrial facilities, astronomy of neutron stars can also throw light on the low T and high n_B region of the QCD phase diagram.

Figure 3 shows the lattice results for the QCD equation of state (EoS) at vanishing chemical potential in the temperature range $100\ MeV \lesssim T \lesssim 1000$ MeV for physical light and strange quark

[2]Unlike a system in *global* equilibrium, here temperature and chemical potential may depend on space-time coordinates.

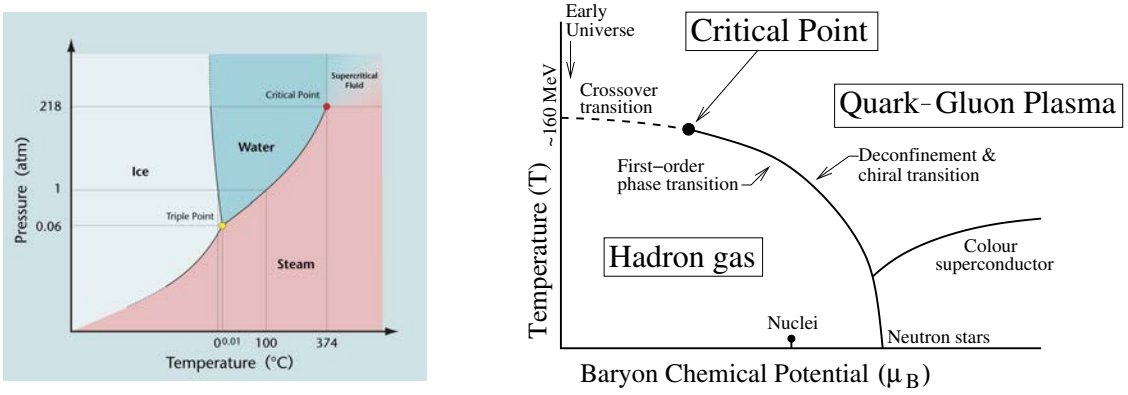


Fig. 2: (a) Phase diagram of water [13] and (b) QCD phase diagram

masses $m_{u,d,s}$. Note that both energy density (ϵ) and pressure (P) rise rapidly around $T = 160$ MeV, indicating an increase in entropy or the number of degrees of freedom. This is consistent with the deconfinement transition with a concomitant release of the partonic degrees of freedom. The rise of P is less rapid than that of ϵ as expected: the square of the speed of sound $c_s^2 = \partial P/\partial\epsilon$ cannot exceed unity. Note also that in the limit of high T, the EoS approaches the form $\epsilon = 3P$ expected of massless particles. However, ϵ is significantly less than ϵ_{SB} showing that the system is far from being in an ideal gaseous state. Lattice results indicate that the transition at vanishing μ_B is merely an analytic crossover. Although there is no strict phase transition, it is common to use the words confined and deconfined phases to describe the low- and high-temperature regimes. For a recent review of the lattice QCD at non-zero temperature, see [22].

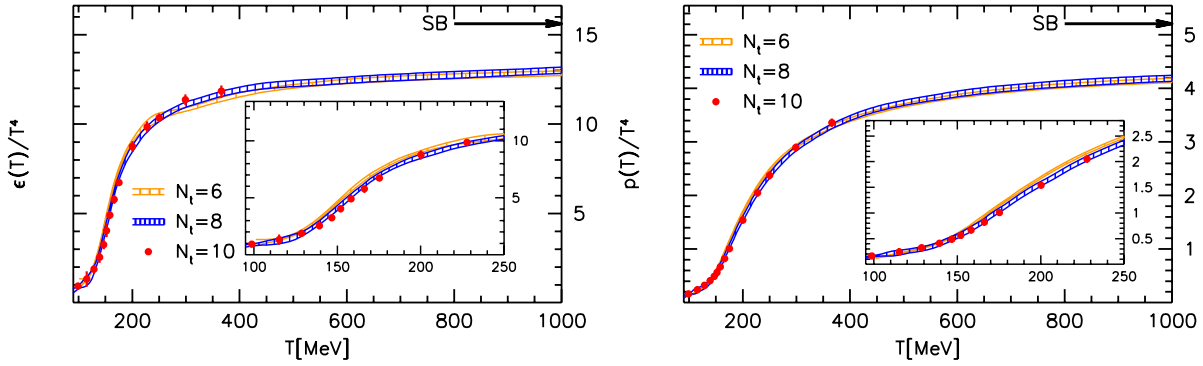


Fig. 3: Energy density and pressure normalized by T^4 as a function of temperature (T) on N_t = 6, 8 and 10 lattices. N_t is the number of lattice points in the temporal direction. The Stefan-Boltzmann (SB) limits are indicated by arrows. Figure from [14]; see also [15].

An ultrarelativistic heavy-ion collision (URHIC) of two (identical) Lorentz-contracted[3] nuclei is thought to proceed as follows. Each incoming nucleus can be looked upon as a coherent [5] cloud of partons (more precisely, a colour-glass-condensate (CGC) plate [24]). The collision results in shattering of the two CGC plates. A significant fraction of the incoming kinetic energy is deposited in the central region leading to a high-energy-density fireball (more precisely, a highly non-equilibrium state called glasma [24]). This is still a coherent state and liberation of partons from the glasma takes a finite amount of (proper) time (a fraction of a fm/c). Subsequently collisions among partons lead to a nearly thermalized (local thermalization!) state called QGP. This happens at a time of the order of 1 fm/c — a

[3]No matter how high the incoming kinetic energy and hence the Lorentz contraction factor is, the limiting thickness of the nucleus is $\sim$ 1 fm due to the so-called wee partons [23].

less understood aspect of the entire process. Due to near thermalization, the subsequent evolution of the system proceeds as per relativistic imperfect fluid dynamics. This involves expansion, cooling, and dilution. Eventually the system hadronizes. Hadrons continue to collide among themselves elastically which changes their energy-momenta, as well as inelastically which alters abundances of individual species. Chemical freezeout occurs when inelastic processes stop. Kinetic freezeout occurs when elastic scatterings too stop. These late stages of evolution when the system is no longer in local equilibrium are simulated using the relativistic kinetic theory framework. Hadrons decouple from the system approximately 10-15 fm$/c$ after the collision and travel towards the surrounding detectors. From the volume of experimental data thus collected one has to establish whether QGP was formed and if so, extract its properties.

After years of work a Standard Model of URHICs has emerged: The initial state is constructed using either the Glauber model [25] or one of the models implementing ideas originating from CGC [26]; for a recent review see [27]. The intermediate evolution is considered using some version of the Müller-Israel-Stewart-like theory [28, 29] of causal relativistic imperfect fluid dynamics, together with a QCD equation of state spanning partonic and hadronic phases [30]. The end evolution of the hadron-rich medium leading to a freezeout uses the Boltzmann equation in the relativistic transport theory [31]. The final state consists of thousands of particles (mesons, baryons, leptons, photons, light nuclei). Detailed measurements (single-particle inclusive, two- and multi-particle correlations, etc.) are available, spanning the energy range from SPS to RHIC to LHC, for various colliding nuclei, centralities, (pseudo)rapidities, and transverse momenta. The aim is to achieve a quantitative understanding of the thermodynamic and transport properties of QGP, e.g., its EoS, its transport coefficients (shear and bulk viscosities, diffusivity, conductivity), etc. The major hurdles in this endeavour are an inadequate knowledge of the initial state and event-to-event fluctuations at nucleonic and sub-nucleonic levels in the initial state.

2 Two most important observables

Elliptic flow and jet quenching are arguably the two most important observables in this field. Observation of an elliptic flow almost as large as that predicted by ideal (i.e., equilibrium) hydrodynamics led to the claim of formation of an almost perfect fluid at RHIC [32]. A natural explanation of the observed jet quenching is in terms of a dense and coloured (hence partonic, not hadronic) medium that is rather opaque to high-momentum hadrons. Recall the definition of QGP given in section 1. The two essential requirements mentioned there seem to be fulfilled considering these two observations together.

Before I discuss these two observations in detail, let me explain what is meant by an almost perfect fluid. Air and water are the two most common fluids we encounter. Which of them is more viscous? Water has a higher coefficient of shear viscosity (η) than air, and appears more viscous. But that is misleading. To compare different fluids, one should consider their kinematic viscosities defined as η/ρ where ρ is the density. Air has a higher kinematic viscosity and hence is actually more viscous than water! Relativistic analogue of η/ρ is the dimensionless ratio η/s where s is the entropy density. Scaling by s is appropriate because number density is ill-defined in the relativistic case. Figure 4 shows constant-pressure ($P_{critical}$) curves for η/s as a function of temperature for various fluids, namely water, nitrogen, helium, and the fluid formed at RHIC. All fluids show a minimum at the critical temperature, and among them the RHIC fluid has the lowest η/s, even lower than that of helium. Hence it is the most perfect fluid observed so far[4]. For water, nitrogen, and helium, points to the left (right) of the minimum refer to the liquid (gaseous) phase. As T rises, η/s for these liquids drops, attains a minimum at the critical temperature T_0, and then in the gaseous phase it rises. This is because liquids and gases transport momentum differently [35]. RHIC fluid is an example of a strongly coupled quantum fluid and has been called sQGP to distinguish it from weakly coupled QGP or wQGP expected at extremely high

[4]More recently, trapped ultracold atomic systems are also shown to have η/s much smaller than that for helium [34].

temperatures. Interestingly, the liquid formed at RHIC and LHC cools into a (hadron resonance) gas!

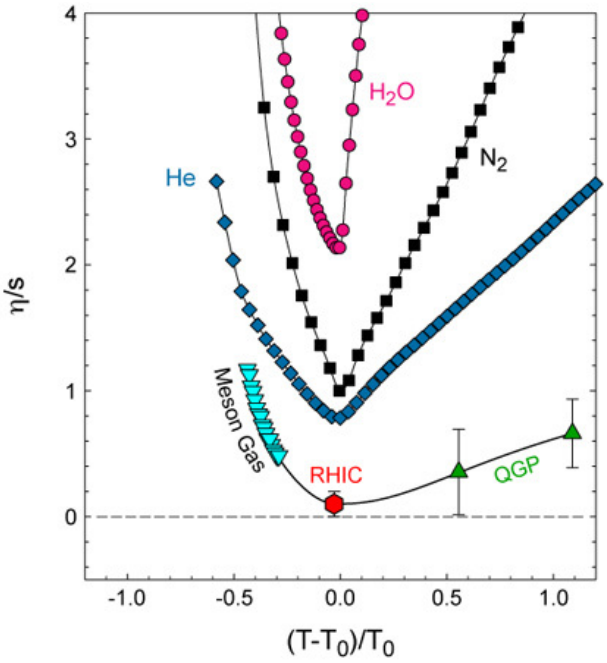


Fig. 4: Constant pressure ($P_{critical}$) curves for (shear viscosity/entropy density) vs temperature. T_0 is the critical temperature of the liquid-gas phase transition. Points labelled Meson Gas are based on chiral perturbation theory and have 50% errors (not shown). Points labelled QGP are based on lattice QCD simulations. Figure from [33].

2.1 Elliptic flow

Consider a non-central (i.e., non-zero impact parameter) collision of two identical spherical nuclei travelling in opposite directions; see Fig. 5(a). In an actual experiment the magnitude and orientation of the impact parameter vector fluctuate from event to event (Fig. 5(b)) and are unknown. This initial geometry can potentially affect the distribution of particles in the final state — in particular, in the transverse plane. In order to capture this physics in terms of a few parameters, the triple differential invariant distribution of particles emitted in the final state is Fourier-decomposed as follows [36]

$$E\frac{d^3N}{d^3p} = \frac{d^3N}{p_T dp_T dy d\phi} = \frac{d^2N}{p_T dp_T dy}\frac{1}{2\pi}\left[1 + \sum_{n=1}^{\infty} 2v_n \cos n(\phi - \Phi_R)\right], \tag{1}$$

where p_T is the transverse momentum, y the rapidity, ϕ the azimuthal angle of the outgoing particle momentum, and Φ_R the reaction-plane angle. Sine terms, $\sin n(\phi - \Phi_R)$, are not included in the Fourier expansion in Eq. (1) because they vanish due to the reflection symmetry with respect to the reaction plane; see Fig. 5. The reaction-plane angle Φ_R which characterizes the initial geometry (Fig. 5(b)) is not known, and is estimated using the transverse distribution of particles in the final state. The estimated reaction plane is called the event plane. The leading term in the square brackets in Eq. (1) represents the azimuthally symmetric radial flow. The first two harmonic coefficients v_1 and v_2 are called directed and elliptic flows, respectively[5]. We have

$$v_n(p_T, y) = \langle \cos[n(\phi - \Phi_R)] \rangle = \frac{\int_0^{2\pi} d\phi \cos[n(\phi - \Phi_R)] \frac{d^3N}{p_T dp_T dy d\phi}}{\int_0^{2\pi} d\phi \frac{d^3N}{p_T dp_T dy d\phi}}. \tag{2}$$

[5]To understand this nomenclature, make polar plots of $r = (1 + 2v_n \cos n\phi)$ for a small positive value of v_n.

The average is taken in the (p_T, y) bin under consideration. After taking the average over all particles in an event, average is then taken over all events in a centrality class[6]. For a central collision the azimuthal distribution is isotropic, and hence $v_n = 0$, i.e., only the radial flow survives. For a review of the methods used for analyzing anisotropic flow in relativistic heavy-ion collisions, and interpretations and uncertainties in the measurements, see [37, 38].

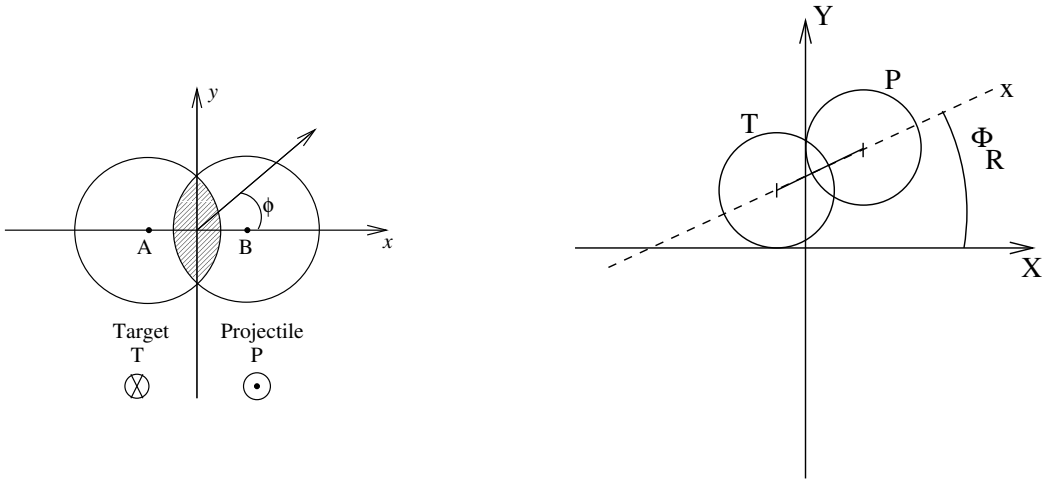


Fig. 5: (a) Non-central collision of two nuclei. Collision or beam axis is perpendicular to the plane of the figure. Impact parameter b = length AB. z is the longitudinal direction, xy is the transverse or azimuthal plane, xz is the reaction plane, and ϕ is the azimuthal angle of one of the outgoing particles. The shaded area indicates the overlap zone. For a central or head-on collision ($b = 0$) the reaction plane cannot be defined. (b) XYZ are the lab-fixed axes. Φ_R is the reaction-plane angle.

In a non-central collision, the initial state is characterized by a *spatial anisotropy* in the azimuthal plane (Fig. 5). Consider particles in the almond-shaped overlap zone. Their initial momenta are predominantly longitudinal. Transverse momenta, if any, are distributed isotropically. If these particles do not interact with each other, the final (azimuthal) distribution too will be isotropic. On the other hand, if they do interact with each other frequently and with adequate strength (or cross section), then the (local) thermal equilibrium is likely to be reached. Once that happens, the system can be described in terms of thermodynamic quantities such as temperature, pressure, etc. The spatial anisotropy of the overlap zone ensures anisotropic pressure gradients in the transverse plane. This leads to a final state characterized by *momentum anisotropy*, an anisotropic azimuthal distribution of particles, and hence a nonvanishing v_n. Thus v_n is a measure of the degree of thermalization of the quark-gluon matter produced in a noncentral heavy-ion collision — a central issue in this field.

The anisotropic flow v_n is sensitive to the *early* ($\sim$ fm/c) history of the collision: Higher pressure gradients along the minor axis of the spatially anisotropic source (Fig. 5) imply that the expansion of the source would gradually diminish its anisotropy, making the flow self-quenching. Thus v_n builds up early (i.e., when the anisotropy is significant) and tends to saturate as the anisotropy continues to decrease. (This is unlike the radial flow which continues to grow until freezeout and is sensitive to early- as well as late-time history of the collision). Thus v_n is a signature of pressure at early times.

The flow v_n depends on the initial conditions, i.e., the beam energy, the mass number of colliding nuclei, and the centrality of the collision. It also depends on the species of the particles under consideration apart from their transverse momentum (p_T) and rapidity (y) or pseudorapidity (η). Using the symmetry of the initial geometry, one can show that $v_n(y)$ is an even (odd) function of y if n is even (odd). Hence $v_1(y)$ vanishes at mid-rapidity. At RHIC energies at mid-rapidity, it is the elliptic flow

[6]Centrality of a AA collision is determined making use of its tight correlation with the charged-particle multiplicity or transverse energy at mid-rapidity, which in turn are anti-correlated with the energy deposited in the Zero Degree Calorimeters.

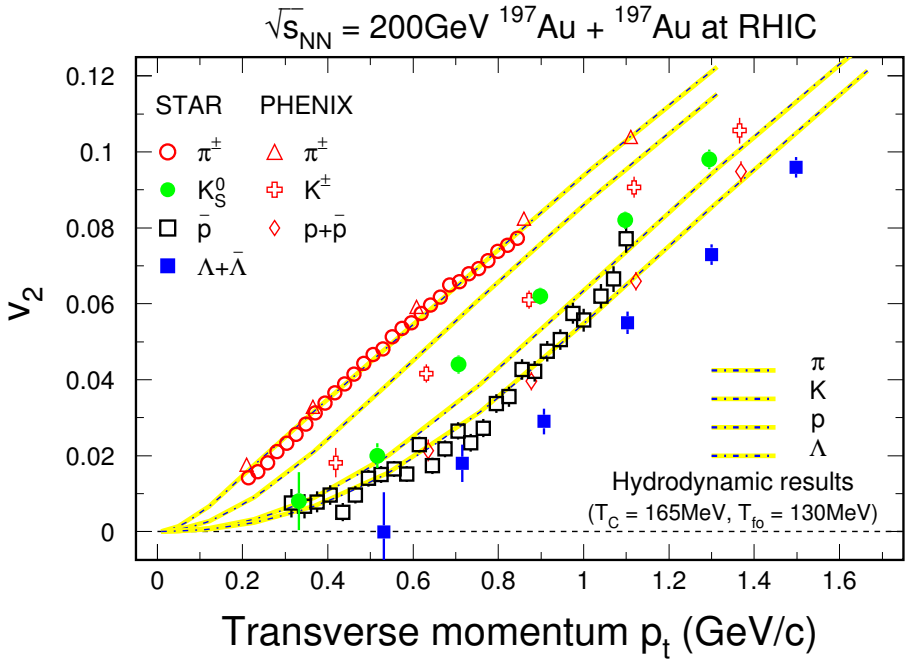


Fig. 6: Success of ideal hydrodynamics: Minimum-bias elliptic flow data for different particle species in comparison with ideal hydrodynamics calculations. Figure from [39].

v_2 that plays an important role. Figure 6 shows the $v_2(p_T)$ data at the highest RHIC energy for various particle species, in broad agreement with the ideal hydrodynamic calculations. As stated before, this success of the ideal hydrodynamics led to the claim of formation of an almost perfect fluid at RHIC.

Extraction of η/s: Introduction of shear viscosity tends to reduce the elliptic flow, v_2, with respect to that for an ideal fluid: a particle moving in the reaction plane (Fig. 5(a)) being faster experiences a greater frictional force compared with a particle moving out of the plane thereby reducing the azimuthal anisotropy and hence v_2. This fact has been used to place an upper limit on the value of η/s of the RHIC fluid. A more precise determination is hindered by ambiguities in the knowledge of the initial state. Event-to-event fluctuations give rise to 'new' flows and observables which help constrain the η/s further.

2.1.1 Event-to-event fluctuations

The discussion above was somewhat idealistic because we assumed smooth initial geometry: Energy (or entropy) density $\epsilon(x, y)$ (or $s(x, y)$) in the shaded area in Fig. 5(a) was a smooth function of x, y because it was assumed to result from the overlap of two smooth Woods-Saxon nuclear density distributions. However, the reality is not so simple, i.e., the initial geometry is not smooth.

In relativistic heavy-ion collisions, the collision time-scale is so short that each incoming nucleus sees nucleons in the other nucleus in a frozen configuration. Event-to-event fluctuations in nucleon (N) positions (and hence in NN collision points) result in an overlap zone with inhomogeneous energy density and a shape that fluctuates from event to event, Fig. 7. This necessitates that the "sine terms" are

also included in the Fourier expansion in Eq. (1). Equivalently, one writes

$$E\frac{d^3N}{d^3p} = \frac{d^3N}{p_T dp_T dy d\phi} = \frac{d^2N}{p_T dp_T dy}\frac{1}{2\pi}\left[1 + \sum_{n=1}^{\infty} 2v_n \cos n(\phi - \Psi_n)\right]. \quad (3)$$

Thus each harmonic n may have its own reference angle Ψ_n in the transverse plane. Traditional hydrodynamic calculations do not take these event-to-event fluctuations into account. Instead of averaging

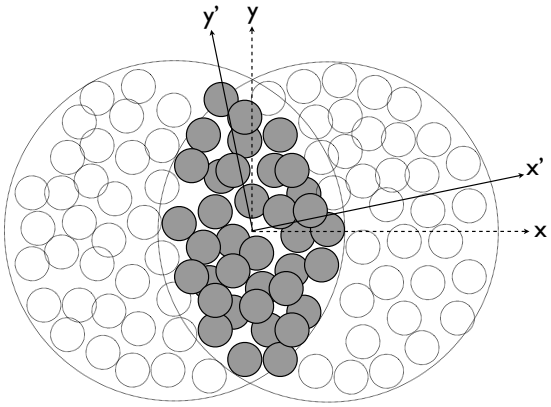


Fig. 7: 'Snapshot' of nucleon positions at the instant of collision. Due to event-to-event fluctuations, the overlap zone could be shifted and tilted with respect to the (x, y) frame. $x'y'$: principal axes of inertia. Figure from [40].

over the fluctuating initial conditions and then evolving the resultant smooth distribution, one needs to perform event-to-event hydrodynamics calculations first and then average over all outputs. This is done in some of the recent hydrodynamic calculations. They also incorporate event-to-event fluctuations at the sub-nucleonic level. Fluctuating initial geometry results in 'new' (rapidity-even) flows (Fig. 8). The rapidity-even dipolar flow shown in Fig. 8(a) is not to be confused with the rapidity-odd directed flow $v_1(p_T, y)$ resulting from the smooth initial geometry in Fig. 5.

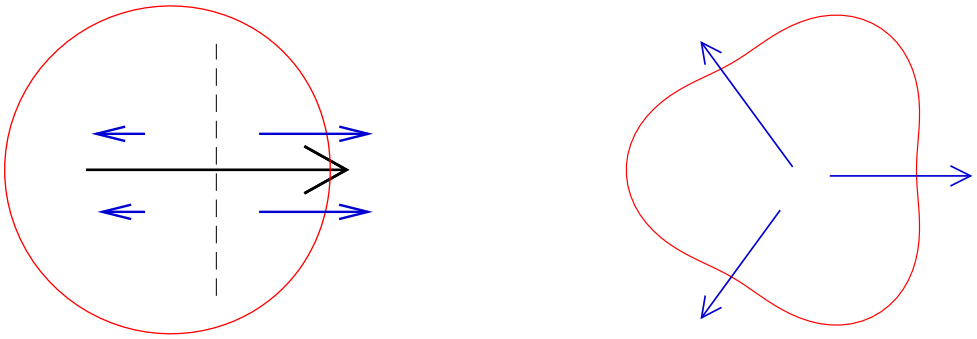


Fig. 8: (a) Dipole asymmetry giving rise to a dipolar flow $v_1(p_T, y)$. The cross indicates the centre of entropy (analogous to the centre of mass) and the large arrow indicates the orientation of the dipole. (b)Triangularity giving rise to a triangular flow $v_3(p_T, y)$. Figure from [41].

For recent reviews of the collective flow, its anisotropies, its event-to-event fluctuations, and the extraction of the specific shear viscosity η/s of QGP, see [42–44].

2.2 Jet quenching

Recall the role played by successively higher-energy electron beams, over many decades in the last century, to unravel the structure of atoms, nuclei, and protons. Studying the properties of QGP by means

of an external probe is obviously ruled out because of its short ($\sim 10^{-23}$ s) life-time. Instead one uses a hard parton produced internally during the nucleus-nucleus collision to probe the medium in which it is produced. Consider, e.g., $g + g \rightarrow g + g$ where two longitudinally moving energetic gluons from the colliding nuclei interact and produce two gluons at large transverse momenta, which fragment and emerge as jets of particles. Hard partons are produced early in the collision: $t \sim 1/Q \sim 1/p_T$, where Q is the parton virtuality scale, and hence they probe the early stages of the collision. Moreover, their production rate is calculable in perturbative QCD. Parton/jet interacts with the medium and loses energy or gets quenched as it traverses the medium (Fig. 9(a)). The amount of energy loss depends among other things on the path length (L) the jet has to travel inside the medium. Figure 9(b) shows the data on the nuclear modification factor, R_{AA}, defined schematically as

$$R_{AA}(p_T) = \text{Yield in } AA / \langle N_{coll} \rangle \text{ Yield in } pp, \tag{4}$$

where $\langle N_{coll} \rangle$ is the mean number of nucleon-nucleon collisions occurring in a single nucleus-nucleus (AA) collision, obtained within the Glauber model [25]. If the nucleus-nucleus collision were a simple superposition of nucleon-nucleon collisions, the ratio R_{AA} would be unity. Direct-photon production rate is consistent with the next-to-leading-order (NLO) perturbative QCD (pQCD) calculation and there is no suppression of the photon yield. However, the yields of high-p_T pions and etas are suppressed by a factor of ~ 5. No such suppression was seen in dAu and pPb collisions [49] (where QGP is not expected to be formed) thereby ruling out suppression by cold nuclear matter as the cause. These observations indicate that the hard-scattered partons lose energy as they traverse the hot medium and the suppression is thus a final-state effect.

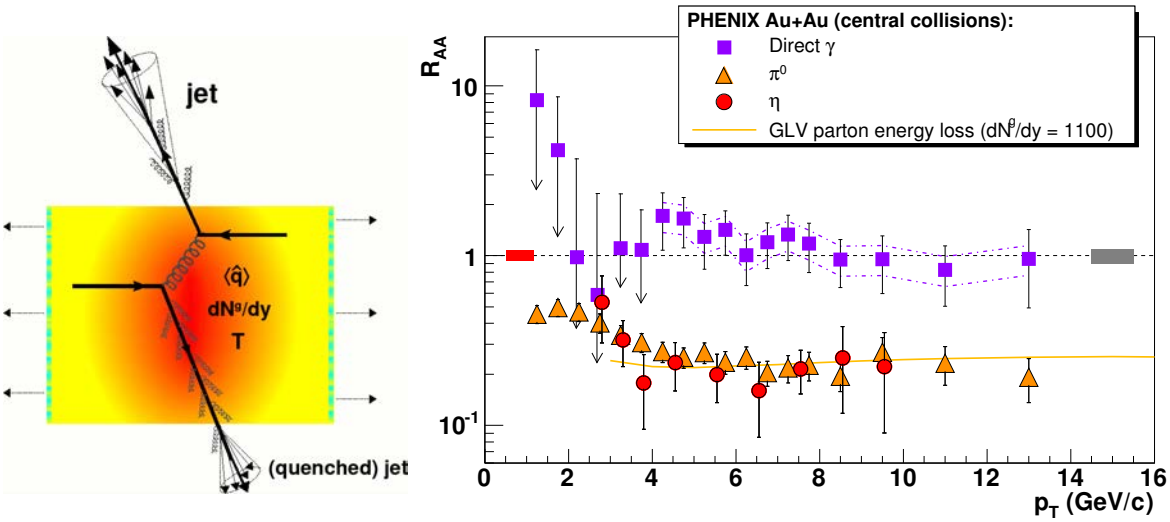


Fig. 9: (a) Back-to-back jets, one produced near the surface of the hot and dense medium and the other deep inside. These are called the near-side and away-side jets. The latter gets quenched. The medium is characterized by its temperature (T), gluon number density in the rapidity space (dN^g/dy), and the transport coefficient or jet-quenching parameter ($\hat{q}$). Figure from [45]. (b) AuAu central collision data on nuclear modification factor R_{AA} as a function of p_T, at the centre-of-mass energy $\sqrt{s_{NN}} = 200$ GeV. Dash-dotted lines: theoretical uncertainties in the direct photon R_{AA}. Solid yellow line: jet-quenching calculation of [46, 47] for leading pions in a medium with initial effective gluon density $dN^g/dy = 1100$. Error bands at $R_{AA} = 1$ indicate the absolute normalization errors. Figure from [48].

Figure 9(b) illustrated jet quenching in a single-particle inclusive yield. Jet quenching is also seen in dihadron angular correlations shown in Fig. 10 as a function of the opening angle between the trigger and associated particles. The only difference between the left and the right panels is the definition of the associated particles. The left panel shows the suppression of the away-side jet in AuAu central, but

not in pp and dAu central collisions. This is expected because unlike AuAu collisions, no hot and dense medium is likely to be formed in pp and dAu collisions, and so there is no quenching of the away-side jet. Energy of the away-side parton in a AuAu collision is dissipated in the medium thereby producing low-p_T or soft particles. When even the soft particles are included, the away-side jet reappears in the AuAu data as shown in the right panel. Its shape is broadened due to interactions with the medium.

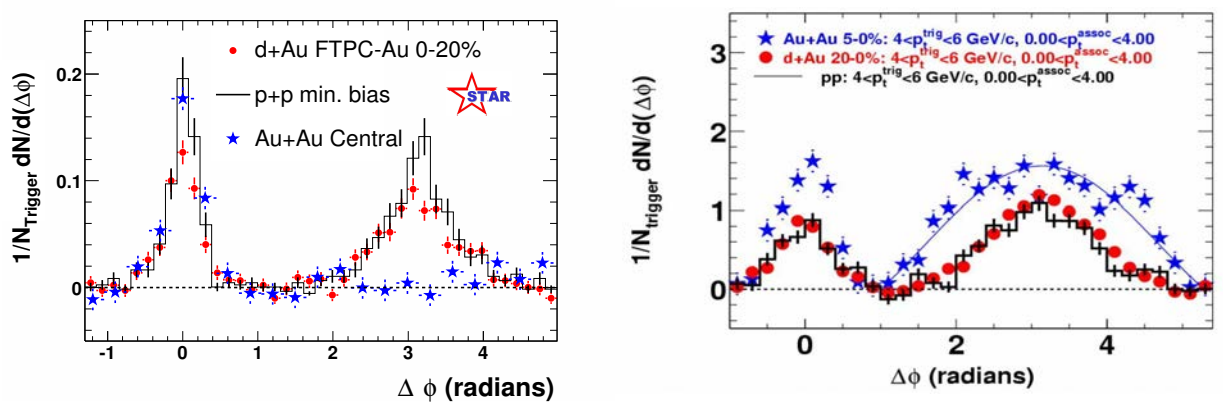


Fig. 10: (a) STAR data on dihadron angular correlations. $\Delta\phi$ is the opening angle between the trigger ($4 < p_T^{trig} < 6$ GeV/c) and associated particles ($2 < p_T^{assoc} < p_T^{trig}$ GeV/c). Figure from [50]. (b) Similar to the left panel, except that $0 < p_T^{assoc} < 4$ GeV/c. Figure from [51].

Figure 11 shows two main mechanisms by which a parton moving in the medium loses energy. Collisional energy loss via elastic scatterings dominates at low momenta whereas the radiative energy loss via inelastic scatterings dominates at high momenta. Energy loss per unit path length depends on the properties of the parton (parton species, energy E), as well as the properties of the medium (T, dN^g/dy, $\hat{q}$). The jet quenching parameter, $\hat{q}$, is defined as the average p_T^2 transferred to the outgoing parton per unit path length. The value of $\hat{q}$ estimated in leading-order QCD is $\simeq 2.2$ GeV^2/fm, while the value extracted from phenomenological fits to the RHIC experimental data on parton energy loss is $\mathcal{O}(10)$ GeV^2/fm.

Fig. 11: Collisional (left) and medium-induced radiative (right) energy loss mechanisms. Their predictions for the energy loss per unit length differ from each other: $\Delta E \propto L$ and $\Delta E \propto L^2$, respectively. Figure from [45].

Jets are more abundant and easier to reconstruct at LHC than at RHIC. Figure 12 shows an example of an unbalanced dijet in a PbPb collision event at CMS (LHC). By studying the evolution of the dijet imbalance as a function of collision centrality and energy of the leading jet, one hopes to get an insight into the dynamics of the jet quenching.

For recent reviews of jet quenching, see e.g., [45, 53–55].

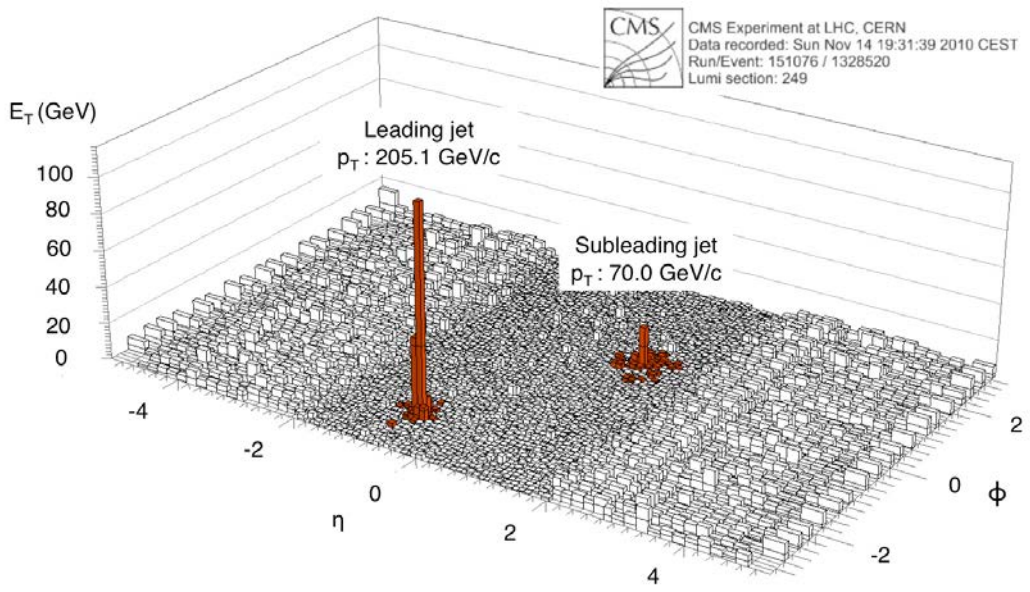


Fig. 12: Jet quenching in PbPb collision at the centre-of-mass energy $\sqrt{s_{NN}} = 2.76$ TeV at CMS. E_T is the summed transverse energy in the electromagnetic and hadron calorimeters. η and ϕ are the pseudorapidity and azimuthal angle, respectively. Figure from [52].

3 Some other important observables

Elliptic flow, or more generally anisotropic collective flow, and jet quenching, which we discussed above are examples of soft and hard probes, respectively. Here 'soft' refers to the low-p_T regime: $0 \lesssim p_T \lesssim 1.5$ GeV/c, and hard refers to high-p_T regime: $p_T \gg 5$ GeV/c. (At RHIC, such high-p_T jets are rare, which explains the relatively low p_T cuts used in Fig. 10.) The medium-p_T regime ($1.5 \lesssim p_T \lesssim 5$ GeV/c) is also interesting, e.g., for the phenomenon of constituent quark number scaling or quark coalescence. In this section we discuss briefly this and other important observables. We shall, however, not discuss a few other important topics such as femtoscopy with two-particle correlation measurements [56–58] and electromagnetic probes of QGP [59, 60].

3.1 Constituent quark number scaling

In the high-p_T regime, hadronization occurs by fragmentation, whereas in the medium-p_T regime, it is modelled by quark recombination or coalescence. The phenomenon of constituent quark number scaling provides experimental support to this model. Figure 13 explains the meaning of constituent quark number (n_q) scaling. In the left panel one sees two distinct branches, one for baryons ($n_q = 3$) and the other for mesons ($n_q = 2$). When scaled by n_q (right panel), the two curves merge into one universal curve, suggesting that the flow is developed at the quark level, and hadrons form by the merging of constituent quarks. This observation provides the most direct evidence for deconfinement so far. ALICE (LHC) has also reported results for the elliptic flow $v_2(p_T)$ of identified particles produced in PbPb collisions at 2.76 TeV. The constituent quark number scaling was found to be not as good as at RHIC [62].

For a recent review see [63].

3.2 Ratios of particle abundances[7]

Ratios of particle abundances such as K/π, p/π, etc. constrain models of particle production. In the thermal or statistical hadronization model [64,65], particles in the final state are assumed to be emitted by

[7] See also section 6.2.2.

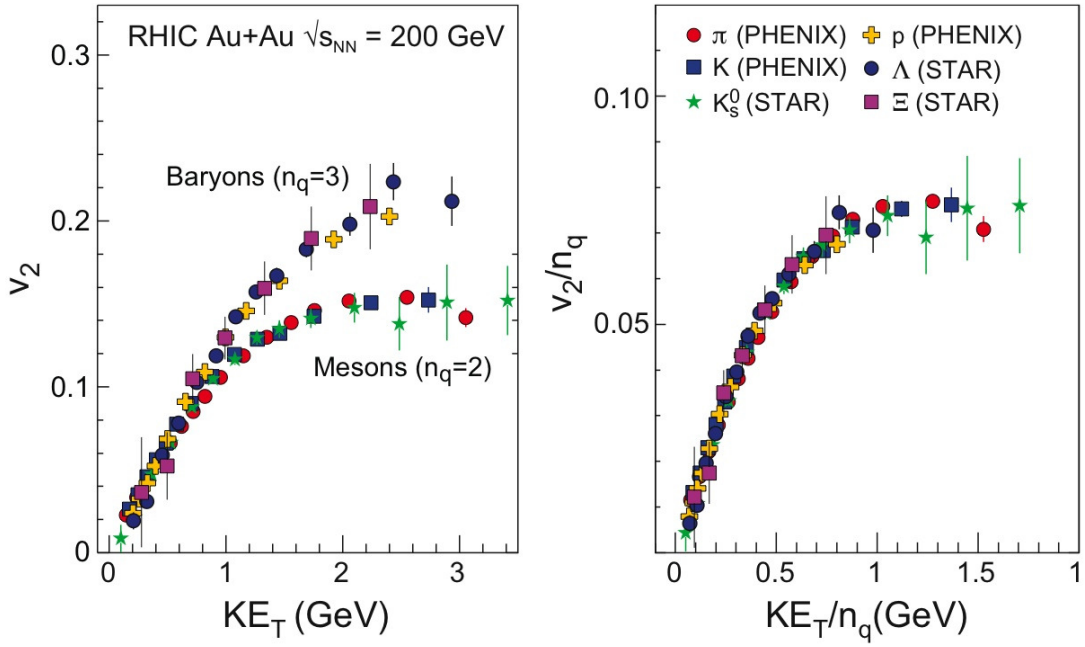


Fig. 13: (Left) Elliptic flow v_2 vs transverse kinetic energy KE_T for various baryons and mesons. (Right) Both v_2 and KE_T are scaled by the number of constituent quarks n_q. Figure from [61].

a source in a thermodynamic equilibrium characterized by only a few parameters such as the (chemical freezeout) temperature and the baryo-chemical potential. These parameters are determined by fitting the experimental data on particle abundances. This model has been quite successful in explaining the Alternating Gradient Synchrotron (AGS), Super Proton Synchrotron (SPS), and RHIC data on the particle ratios [66, 67]. These facilities together cover the centre-of-mass energy ($\sqrt{s_{NN}}$) range from 2 GeV to 200 GeV.

For a recent review of the statistical hadronization picture with an emphasis on charmonium production, see [68].

3.3 Strangeness enhancement

Production of strange particles is expected to be enhanced [69, 70] in relativistic nucleus-nucleus collisions relative to the scaled up pp data (Eq. (4)) because of the following reasons: (1) Although $m_s \gg m_{u,d}$, strange quarks and antiquarks can be abundant in an equilibrated QGP with temperature $T > m_s$, (2) large gluon density in QGP leads to an efficient production of strangeness via gluon fusion $gg \rightarrow s\bar{s}$, and (3) energy threshold for strangeness production in the purely hadron-gas scenario is much higher than in QGP. Abundance of strange quarks and antiquarks in QGP is expected to leave its imprint on the number of strange and multi-strange hadrons detected in the final state. The above expectation was borne out by the measurements made at SPS and RHIC; see Fig. 14 where N_{part} is the mean number of participating nucleons in a nucleus-nucleus collision, estimated using the Glauber model [25] and serves as a measure of the centrality of the collision. The idea of strangeness enhancement in AA collisions or equivalently of strangeness suppression in pp collisions can be recast in the language of statistical mechanics of grand canonical (for central AA collisions) and canonical (for pp collisions) ensembles; see, e.g., [71]. A complete theoretical understanding of these results is yet to be achieved [71].

For a review of strange hadron production in heavy-ion collisions from SPS to RHIC, see [72]. For the ALICE (LHC) results on multi-strange baryon production at 2.76 TeV, see [73]. ALICE observed that the strangeness enhancement was less pronounced than at lower energies.

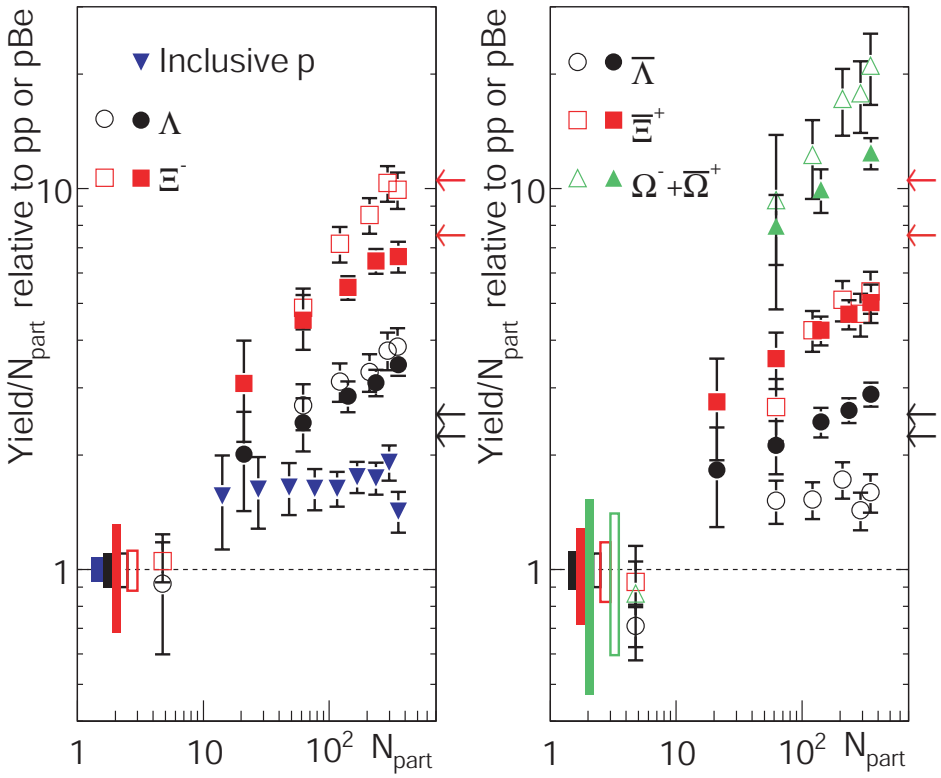


Fig. 14: Enhanced strange baryon production as a function of $\langle N_{part} \rangle$, at mid-rapidity, in AA collisions compared to $\langle N_{part} \rangle$-scaled pp interactions at the same energy. Solid markers: STAR data on AuAu collisions at $\sqrt{s_{NN}} =$ 200 GeV. Open symbols: SPS data on PbPb collisions at $\sqrt{s_{NN}} = 17.3$ GeV. Boxes at unity show statistical and systematic uncertainties and arrows on the right axes mark the predictions of a thermal model. Figure from [71].

3.4 Sequential melting of heavy quarkonia[8]

Colour Debye screening of the attraction between heavy quarks (c or b) and antiquarks ($\bar{c}$ or $\bar{b}$) in a hot and dense medium such as QGP is expected to suppress the formation of quarkonia relative to what one expects from a pp baseline measurement [74]. Observation of suppression would thus serve as a signal for deconfinement. As the temperature of the medium rises, various quarkonium states are expected to 'melt' one by one in the sequence of their increasing binding energies. The sequential melting of heavy quarkonia thus serves as a 'thermometer' for the medium. A reliable estimation of the charmonium[9] formation rates, however, needs to take into account several other competing effects:

- gluon shadowing/anti-shadowing and saturation effects in the initial wave functions of the colliding nuclei,
- initial- and final-state k_T scatterings and parton-energy loss,
- charmonium formation via colour-singlet and colour-octet channels,
- feed-down from the excited states of the charmonium to its ground state,
- secondary charmonium production by recombination or coalescence of independently produced c and $\bar{c}$,
- interaction of the outgoing charmonium with the medium, etc.

A systematic study of suppression patterns of J/ψ and Υ families, together with pA baseline measurements, over a broad energy range, would help disentangle these hot and cold nuclear matter effects.

[8] See also section 6.2.3.
[9] Similar statements would be true for the bottomonium.

Table 1: Big Bang and Little Bang comparison

	Big Bang	Little Bang
Occurrence	Only once	Millions of times at RHIC, LHC
Initial state	Inflation? (10^{-35} s)	Glasma? (10^{-24} s)
Expansion	General Relativity	Rel. imperfect fluid dynamics
Freezeout temperatures	γ : 2.73 K, ν : 1.95 K	$T_{ch} \sim 150, T_{kin} \sim 120$ MeV
Anisotropy in	Final temp. (CMB)	Final flow profile
Penetrating probes	Photons	Photons, jets
Chemical probes	Light nuclei	Various hadron species
Colour shift	Red shift	Blue shift
Tools	COBE, WMAP, Planck	SPS, RHIC, LHC
Starting years	1989, 2001, 2009	1987, 2000, 2009

For reviews of charmonium and/or bottomonium production in heavy-ion collisions, see [75–78]. For a review of heavy-flavour probes of the QCD matter formed at RHIC, see [79].

4 Big Bang and Little Bang

Having described the various stages in the relativistic heavy-ion collisions and the most important observables and probes in this field, let me bring out the striking similarities between the Big Bang and the Little Bang. In both cosmology and the physics of relativistic heavy-ion collisions, the initial quantum fluctuations ultimately lead to macroscopic fluctuations and anisotropies in the final state. In both the fields, the goal is to learn about the early state of the matter from the final-state observations. See Table 1 for the comparison of these two fields. Here 2.73 K and 1.95 K are photon and neutrino decoupling or freezeout temperatures, respectively. T_{ch} and T_{kin} are the chemical and kinetic freezeout temperatures mentioned in section 1. The last two rows list the various experimental 'tools' and the years in which they were commissioned. For a more detailed comparison, see [5, 80, 81].

5 Fluid dynamics

The kinetic or transport theory of gases is a microscopic description in the sense that detailed knowledge of the motion of the constituents is required. Fluid dynamics (also loosely called hydrodynamics) is an effective (macroscopic) theory that describes the slow, long-wavelength motion of a fluid close to local thermal equilibrium. No knowledge of the motion of the constituents is required to describe observable phenomena. Quantitatively, if l denotes the mean free path, τ the mean free time, k the wave number, and ω the frequency, then $kl \ll 1,\ \omega\tau \ll 1$ is the hydrodynamic regime, $kl \simeq 1,\ \omega\tau \simeq 1$ the kinetic regime, and $kl \gg 1,\ \omega\tau \gg 1$ the nearly-free-particle regime.

Relativistic hydrodynamic equations are a set of coupled partial differential equations for number density n, energy density ϵ, pressure P, hydrodynamic four-velocity u^μ, and in the case of imperfect hydrodynamics, also bulk viscous pressure Π, particle-diffusion current n^μ, and shear stress tensor $\pi^{\mu\nu}$. In addition, these equations also contain the coefficients of shear and bulk viscosities and thermal conductivity, and the corresponding relaxation times. Further, the equation of state (EoS) needs to be supplied to make the set of equations complete. Hydrodynamics is a powerful technique: Given the initial conditions and the EoS, it predicts the evolution of the matter. Its limitation is that it is applicable at or near (local) thermal equilibrium only.

Relativistic hydrodynamics finds applications in cosmology, astrophysics, high-energy nuclear physics, etc. In relativistic heavy-ion collisions, it is used to calculate the multiplicity and transverse

momentum spectra of hadrons, anisotropic flows, and femtoscopic radii. Energy density or temperature profiles resulting from the hydrodynamic evolution are needed in the calculations of jet quenching, J/ψ melting, thermal photon and dilepton productions, etc. Thus hydrodynamics plays a central role in modeling relativistic heavy-ion collisions.

Hydrodynamics is formulated as an order-by-order expansion in the sense that in the first (second)-order theory, the equations for the dissipative fluxes contain the first (second) derivatives of u^μ. The ideal hydrodynamics is called the zeroth-order theory. The zeroth-, first-, and second-order equations are named after Euler, Navier-Stokes, and Burnett, respectively, in the non-relativistic case (Fig. 15). The relativistic Navier-Stokes equations are parabolic in nature and exhibit acausal behaviour, which was rectified in the (relativistic second-order) Israel-Stewart (IS) theory [29]. The formulation of the relativistic imperfect second-order hydrodynamics ('2' in Fig. 15) is currently under intense investigation; see, e.g., [82–86] for the recent activity in this area. Hydrodynamics has traditionally been derived either from entropy considerations (i.e., the generalized second law of thermodynamics) or by taking the second moment of the Boltzmann equation.

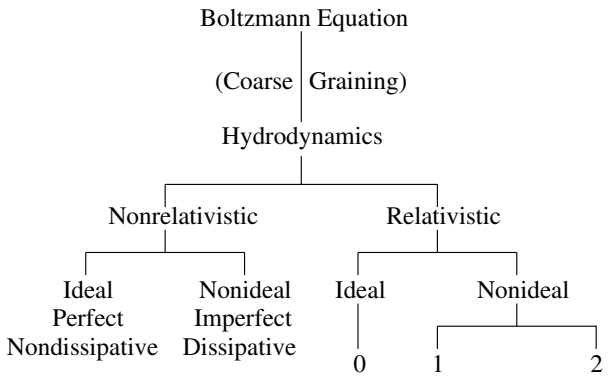


Fig. 15: Coarse-Graining of the Boltzmann equation

For a comprehensive treatment of relativistic hydrodynamics, numerical techniques, and applications, see [87]. For an elementary introduction to relativistic hydrodynamics and its application to heavy-ion collisions, see [88]. For a review of new developments in relativistic viscous hydrodynamics, see [89].

6 LHC highlights

6.1 RHIC-LHC comparison

Table 2 compares some basic results obtained at LHC soon after it started operating, with similar results obtained earlier at RHIC. Here $dN_{ch}/d\eta$ is the charged particle pseudorapidity density, at mid-rapidity, normalized by $\langle N_{part}\rangle/2$ where $\langle N_{part}\rangle$ is the mean number of participating nucleons in a nucleus-nucleus collision, estimated using the Glauber model [25]. ϵ_{Bj} is the initial energy density estimated using the well-known Bjorken formula [5, 7]. τ_i is the initial or formation time of QGP. Assuming conservatively the same $\tau_i \simeq 0.5$ fm at LHC as at RHIC, one gets an estimate of ϵ_{Bj} at LHC. T_i is the initial temperature fitted to reproduce the observed multiplicity of charged particles in a hydrodynamical model. Note that the $\sim 30\%$ increase in T_i is consistent with the factor of ~ 3 rise in ϵ_{Bj}. $V_{f.o.}$ is the volume of the system at the freezeout, measured with two-pion Bose-Einstein correlations. v_{flow} is the radial velocity of the collective flow of matter. v_2 is the elliptic flow. It is clear from Table 2 that the QGP fireball produced at LHC is hotter, larger, and longer-lasting, as compared with that at RHIC.

Table 2: RHIC-LHC comparison

	RHIC (AuAu)	LHC (PbPb)	Increase by factor or %
$\sqrt{s_{NN}}$ (GeV)	200	2760	14
$dN_{ch}/d\eta/\left(\frac{<N_{part}>}{2}\right)$	3.76	8.4	2.2
$\epsilon_{Bj}\tau_i$ (GeV/fm^2)	16/3	16	3
ϵ_{Bj} (GeV/fm^3)	10	30	3
T_i (MeV)	360	470	30%
$V_{f.o.}$ (fm^3)	2500	5000	2
Lifetime (fm/c)	8.4	10.6	26%
v_{flow}	0.6	0.66	10%
$< p_T >_\pi$ (GeV)	0.36	0.45	25%
Differential $v_2(p_T)$			unchanged
p_T-integrated v_2			30%

6.2 Some surprises at LHC

6.2.1 Charged-particle production at LHC

Figure 16 presents perhaps the most basic observable in heavy-ion collisions — the number of charged particles produced. This observable helps place constraints on the particle production mechanisms and provides a first rough estimate of the initial energy density reached in the collision. The left panel compares the charged-particle production in central AA and non-single-diffractive (NSD)[10] $pp(p\bar{p})$ collisions at various energies and facilities. The curves are simple parametric fits to the data; note the higher power of s_{NN} in the former case. The precise magnitude of $dN_{ch}/d\eta$ measured in PbPb collisions at LHC was somewhat on a higher side than expected. Indeed, as is clear from the figure, the logarithmic extrapolation of the lower-energy measurements at AGS, SPS, and RHIC grossly under-predicts the LHC data. The right panel highlights an even more surprising fact that the shape of the plotted observable vs centrality is nearly independent of the centre-of-mass energy, except perhaps for the most peripheral AA collisions. Studying the centrality dependence of the charged-particle production throws light on the roles played by hard scatterings and soft processes. For details, see [91].

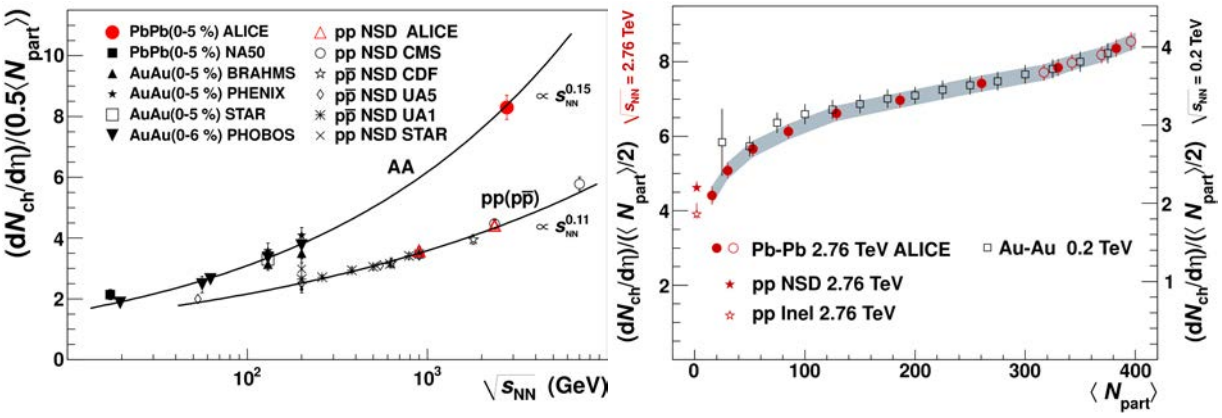


Fig. 16: Charged-particle pseudorapidity density (at $\eta = 0$) per colliding nucleon pair vs $\sqrt{s_{NN}}$ (left panel) and $\langle N_{part}\rangle$ (right panel). Figure from [90].

[10]Non-single-diffractive pp collisions are those which exclude the elastic scattering and single-diffractive events.

6.2.2 Particle ratios at LHC — Proton anomaly

We described above in section 3.2 the success of the thermal/statistical hadronization model in explaining the ratios of particle abundances measured at AGS, SPS, and RHIC. When extended to the LHC energies, however, the model was unable to reproduce the p/π^+ and $\bar{p}/\pi^-$ ratios; the absolute $p, \bar{p}$ yields were off by almost three standard deviations (Fig. 17). Current attempts to understand these discrepancies focus on the possible effects of (a) as yet undiscovered hadrons, or in other words, the incomplete hadron spectrum, (b) the annihilation of some $p, \bar{p}$ in the final hadronic phase, or (c) the out-of-equilibrium physics currently missing in the model. None of these effects has been found to be satisfactory because while reducing the (Data-Fit) discrepancy at one place, it worsens it at other place(s) [92]. Finally, Fig. 17 also shows that most antiparticle/particle ratios are unity within error bars indicating a vanishing baryo-chemical potential at LHC.

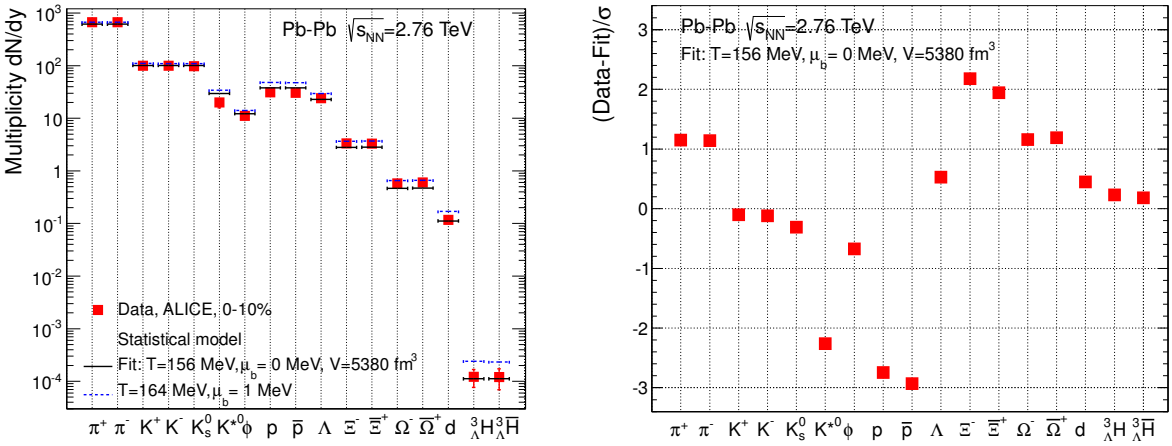


Fig. 17: Left: Hadron yields from ALICE (LHC) together with the fit based on the thermal model (solid black lines). The data point for K^{*0} is not included in the fit. Blue dotted lines show results of the model for the indicated values of T and μ_b, normalized to the value for π^+. Right: Deviations between the thermal fit and the data. Note that the p and $\bar{p}$ yields are below the thermal fit by 2.7 and 2.9 sigma, respectively, whereas the cascade yields are above the fit by about two sigma. Figures from [92].

6.2.3 Quarkonium story at LHC

We described above in section 3.4 the melting of heavy quarkonium as a possible signature of deconfinement or colour screening effects in QGP. Anomalous suppression of J/ψ was first seen at SPS. No significant differences in the suppression pattern were observed at RHIC. LHC, however, has thrown some surprises which are not yet fully understood. Figure 18 presents the nuclear modification factor R_{AA} of J/ψ as a function of centrality (left) and p_T (right), at similar rapidities. Note the differences between the PHENIX and ALICE measurements. Differences at low p_T in the right-hand panel are possibly because of the larger recombination probability at ALICE than at PHENIX; this probability is expected to decrease at high p_T. Sequential suppression of upsilon states was observed by CMS in PbPb collisions at 2.76 TeV: The R_{AA} values for Υ(1S), Υ(2S), and Υ(3S), were about 0.56, 0.12, and lower than 0.10, respectively [94]. For the status of the evolving quarkonium saga, see [95].

7 Concluding remarks

(1) Quark-gluon plasma has been discovered, and we are in the midst of trying to determine its thermodynamic and transport properties accurately.

(2) Data on the collective flow at RHIC/LHC have provided a strong support to hydrodynamics as the

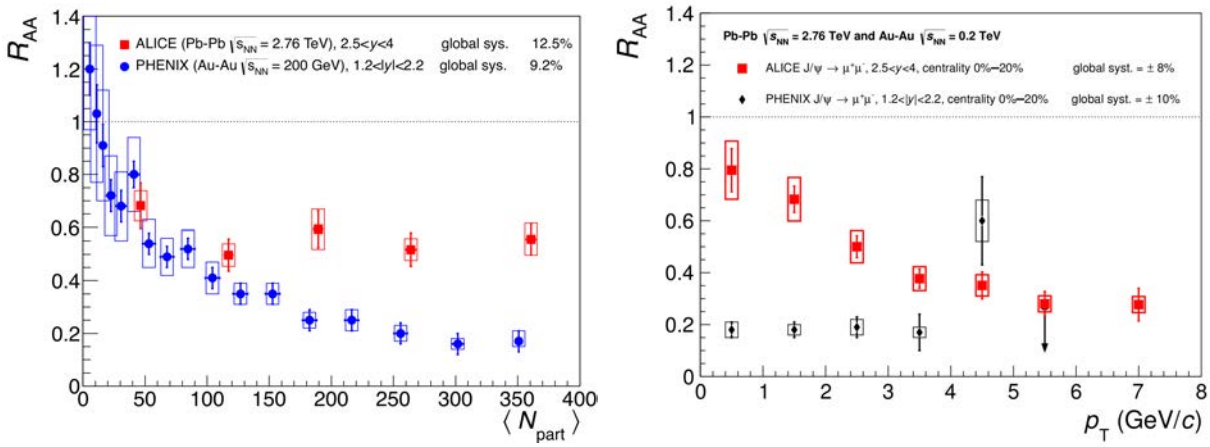


Fig. 18: Nuclear modification factor R_{AA} of J/ψ vs centrality (left) and p_T (right). Figures from [90] and [93].

appropriate effective theory for relativistic heavy-ion collisions. The most complete event-to-event hydrodynamic calculations to date [43, 96] have yielded $\eta/s = 0.12$ and 0.20 at RHIC (AuAu, 200 GeV) and LHC (PbPb, 2.76 TeV), respectively, with at least 50% systematic uncertainties. These are the average values over the temperature histories of the collisions. Uncertainties associated with (mainly) the initial conditions have so far prevented a more precise determination of η/s.

(3) Surprisingly, even the pp collision data at 7 TeV are consistent with the hydrodynamic picture, if the final multiplicity is sufficiently large!

(4) An important open question is at what kinematic scale partons lose their quasiparticle nature (evident in jet quenching) and become fluid like (as seen in the collective flow)?

(5) QCD phase diagram still remains largely unknown.

(6) RHIC remains operational. ALICE, ATLAS, and CMS at LHC all have come up with many new results on heavy-ion collisions. Further updates of these facilities are planned or being proposed. Compressed baryonic matter experiments at FAIR [10] and NICA [97], which will probe the QCD phase diagram in a high baryon density but relatively low temperature region, are a few years in the future. Electron-ion collider (EIC) has been proposed to understand the glue that binds us all [98]. So this exciting field is going to remain very active for a decade at least.

Many review articles have been cited throughout the text above. Here are a few more published in the last 2-3 years [99, 100]. See also these two talks given at the '2013 Nobel Symposium on LHC Physics' for an overview of the status of this field: [101, 102].

Acknowledgements

I sincerely thank Saumen Datta for a critical reading of the manuscript.

References

[1] H. Satz, *Extreme States of Matter in Strong Interaction Physics*, Lect. Notes in Phys. **841** (2012) 1, (Springer, Heidelberg, 2012).

[2] W. Florkowski, *Phenomenology of Ultra-relativistic Heavy-Ion Collisions*, (World Scientific, Singapore, 2010).

[3] J. Bartke, *Relativistic Heavy-Ion Physics*, (World Scientific, Singapore, 2008).

[4] R. Vogt, *Ultra-relativistic Heavy-Ion Collisions*, (Elsevier Science, Amsterdam, 2007).

[5] K. Yagi, T. Hatsuda, and Y. Miake, *Quark-Gluon Plasma*, (Cambridge University Press, Cambridge, 2005).

[6] J. Letessier and J. Rafelski, *Hadrons and Quark-Gluon Plasma*, (Cambridge University Press, Cambridge, 2002).

[7] C.Y. Wong, *Introduction to High-Energy Heavy-Ion Collisions*, (World Scientific, Singapore, 1994).

[8] L. Csernai, *Introduction to Relativistic Heavy-Ion Collisions*, (John Wiley, 1994).

[9] B. Müller, *The Physics of the Quark-Gluon Plasma*, Lect. Notes in Phys. **225** (1985) 1, (Springer, Heidelberg, 1985).

[10] B. Friman, C. Hohne, J. Knoll, S. Leupold, J. Randrup, R. Rapp, and P. Senger, (eds.), *The CBM physics book: Compressed baryonic matter in laboratory experiments*, Lect. Notes in Phys. **814** (2011) 1, (Springer, Heidelberg, 2011).

[11] R.C. Hwa and X.N. Wong, (eds.), *Quark-Gluon Plasma*, Vols. 3, 4, (World Scientific, Singapore, 2004, 2010).

[12] R.C. Hwa, (ed.), *Quark-Gluon Plasma*, Vols. 1, 2, (World Scientific, Singapore, 1990, 1995).

[13] Illustration by Pete Harrison, Langlo Press, reproduced with permission.

[14] S. Borsanyi *et al.*, JHEP **1011** (2010) 077.

[15] S. Borsanyi *et al.*, arXiv:1309.5258 [hep-lat].

[16] K. Rajagopal and F. Wilczek, in M. Shifman (ed.): At the frontier of particle physics, vol. 3, 2061-2151.

[17] M. G. Alford, A. Schmitt, K. Rajagopal, and T. Schaefer, Rev. Mod. Phys. **80** (2008) 1455.

[18] R. Anglani *et al.*, arXiv:1302.4264 [hep-ph].

[19] L. McLerran and R. D. Pisarski, Nucl. Phys. A **796** (2007) 83.

[20] J. Adams *et al.* [STAR Collaboration], Nucl. Phys. A **757** (2005) 102.

[21] K. Rajagopal, Nucl. Phys. A **661** (1999) 150.

[22] P. Petreczky, J. Phys. G **39** (2012) 093002.

[23] J. D. Bjorken, Lect. Notes Phys. **56** (1976) 93.

[24] F. Gelis, E. Iancu, J. Jalilian-Marian, and R. Venugopalan, Ann. Rev. Nucl. Part. Sci. **60** (2010) 463.

[25] M. L. Miller, K. Reygers, S. J. Sanders, and P. Steinberg, Ann. Rev. Nucl. Part. Sci. **57** (2007) 205.

[26] D. Kharzeev and M. Nardi, Phys. Lett. B **507** (2001) 121.

[27] J. L. Albacete and C. Marquet, arXiv:1401.4866 [hep-ph].

[28] I. Muller, Z. Phys. **198** (1967) 329.

[29] W. Israel and J. M. Stewart, Annals Phys. **118** (1979) 341.

[30] P. Huovinen and P. Petreczky, Nucl. Phys. A **837** (2010) 26.

[31] H. Song, S. A. Bass, U. Heinz, T. Hirano, and C. Shen, Phys. Rev. C **83** (2011) 054910 [Erratum-ibid. C **86** (2012) 059903].

[32] M. Gyulassy and L. McLerran, Nucl. Phys. A **750** (2005) 30.

[33] R. A. Lacey *et al.*, Phys. Rev. Lett. **98** (2007) 092301.

[34] T. Schaefer and D. Teaney, Rept. Prog. Phys. **72** (2009) 126001.

[35] L. P. Csernai, J. .I. Kapusta, and L. D. McLerran, Phys. Rev. Lett. **97** (2006) 152303.

[36] S. Voloshin and Y. Zhang, Z. Phys. C **70** (1996) 665.

[37] A. M. Poskanzer and S. A. Voloshin, Phys. Rev. C **58** (1998) 1671.

[38] S. A. Voloshin, A. M. Poskanzer, and R. Snellings, arXiv:0809.2949 [nucl-ex].

[39] M. D. Oldenburg [STAR Collaboration], J. Phys. G **31** (2005) S437.

[40] B. Alver *et al.*, Phys. Rev. C **77** (2008) 014906.
[41] D. Teaney and L. Yan, Phys. Rev. C **83** (2011) 064904.
[42] U. Heinz and R. Snellings, Ann. Rev. Nucl. Part. Sci. **63** (2013) 123.
[43] C. Gale, S. Jeon, and B. Schenke, Int. J. Mod. Phys. A **28** (2013) 1340011.
[44] P. Huovinen, Int. J. Mod. Phys. E **22** (2013) 1330029.
[45] D. d'Enterria, arXiv:0902.2011 [nucl-ex].
[46] I. Vitev and M. Gyulassy, Phys. Rev. Lett. **89** (2002) 252301.
[47] I. Vitev, J. Phys. G **30** (2004) S791.
[48] S. S. Adler *et al.* [PHENIX Collaboration], Phys. Rev. C **75** (2007) 024909.
[49] B. Abelev *et al.* [ALICE Collaboration], Phys. Rev. Lett. **110** (2013) 082302.
[50] J. Adams *et al.* [STAR Collaboration], Phys. Rev. Lett. **91** (2003) 072304.
[51] J. Adams *et al.* [STAR Collaboration], Phys. Rev. Lett. **95** (2005) 152301.
[52] S. Chatrchyan *et al.* [CMS Collaboration], Phys. Rev. C **84** (2011) 024906.
[53] U. A. Wiedemann, arXiv:0908.2306 [hep-ph].
[54] A. Majumder and M. Van Leeuwen, Prog. Part. Nucl. Phys. A **66** (2011) 41.
[55] M. Spousta, Mod. Phys. Lett. A **28** (2013) 1330017.
[56] B. Tomasik and U. A. Wiedemann, In *Hwa, R.C. (ed.) et al.: Quark gluon plasma* 715-777 [hep-ph/0210250].
[57] S. S. Padula, Braz. J. Phys. **35** (2005) 70.
[58] M. A. Lisa, S. Pratt, R. Soltz, and U. Wiedemann, Ann. Rev. Nucl. Part. Sci. **55** (2005) 357.
[59] G. David, R. Rapp, and Z. Xu, Phys. Rept. **462** (2008) 176.
[60] I. Tserruya, arXiv:0903.0415 [nucl-ex].
[61] B. Müller, Acta Phys. Polon. B **38** (2007) 3705.
[62] F. Noferini [ALICE Collaboration], Nucl. Phys. A **904-905** (2013) 483c.
[63] R. J. Fries, V. Greco, and P. Sorensen, Ann. Rev. Nucl. Part. Sci. **58** (2008) 177.
[64] P. Braun-Munzinger, K. Redlich, and J. Stachel, In *Hwa, R.C. (ed.) *et al.*: Quark gluon plasma* 491-599 [nucl-th/0304013].
[65] F. Becattini, arXiv:0901.3643 [hep-ph].
[66] J. Cleymans and K. Redlich, Phys. Rev. Lett. **81** (1998) 5284. [nucl-th/9808030].
[67] A. Andronic, P. Braun-Munzinger, and J. Stachel, Phys. Lett. B **673** (2009) 142 [Erratum-ibid. B **678** (2009) 516].
[68] P. Braun-Munzinger and J. Stachel, arXiv:0901.2500 [nucl-th].
[69] J. Rafelski and B. Müller, Phys. Rev. Lett. **48** (1982) 1066 [Erratum-ibid. **56** (1986) 2334].
[70] P. Koch, B. Müller, and J. Rafelski, Phys. Rept. **142** (1986) 167.
[71] B. I. Abelev *et al.* [STAR Collaboration], Phys. Rev. C **77** (2008) 044908.
[72] C. Blume and C. Markert, Prog. Part. Nucl. Phys. **66** (2011) 834.
[73] B. B. Abelev *et al.* [ALICE Collaboration], Phys. Lett. B **728** (2014) 216.
[74] T. Matsui and H. Satz, Phys. Lett. B **178** (1986) 416.
[75] L. Kluberg and H. Satz, arXiv:0901.3831 [hep-ph].
[76] O. Linnyk, E. L. Bratkovskaya, and W. Cassing, Int. J. Mod. Phys. E **17** (2008) 1367.
[77] R. Rapp, D. Blaschke, and P. Crochet, Prog. Part. Nucl. Phys. **65** (2010) 209.
[78] N. Brambilla, *et al.*, Eur. Phys. J. C **71** (2011) 1534.
[79] A. D. Frawley, T. Ullrich, and R. Vogt, Phys. Rept. **462** (2008) 125.
[80] A. P. Mishra, R. K. Mohapatra, P. S. Saumia, and A. M. Srivastava, Phys. Rev. C **77** (2008) 064902.

[81] A. P. Mishra, R. K. Mohapatra, P. S. Saumia, and A. M. Srivastava, Phys. Rev. C **81** (2010) 034903.
[82] G. S. Denicol, H. Niemi, E. Molnar, and D. H. Rischke, Phys. Rev. D **85** (2012) 114047.
[83] G. S. Denicol, E. MolnÃąr, H. Niemi, and D. H. Rischke, Eur. Phys. J. A **48** (2012) 170.
[84] A. Jaiswal, R. S. Bhalerao, and S. Pal, Phys. Rev. C **87** (2013) 021901 (R).
[85] A. Jaiswal, R. S. Bhalerao, and S. Pal, Phys. Lett. B **720** (2013) 347.
[86] R. S. Bhalerao, A. Jaiswal, S. Pal, and V. Sreekanth, arXiv:1312.1864 [nucl-th].
[87] L. Rezzolla and O. Zanotti, *Relativistic Hydrodynamics*, (Oxford University Press, Oxford, 2013).
[88] J. -Y. Ollitrault, Eur. J. Phys. **29** (2008) 275.
[89] P. Romatschke, Int. J. Mod. Phys. E **19** (2010) 1.
[90] K. Reygers [ALICE Collaboration], arXiv:1208.1626 [nucl-ex].
[91] E. Abbas *et al.* [ALICE Collaboration], Phys. Lett. B **726** (2013) 610.
[92] J. Stachel, A. Andronic, P. Braun-Munzinger, and K. Redlich, arXiv:1311.4662 [nucl-th].
[93] B. B. Abelev *et al.* [ALICE Collaboration], arXiv:1311.0214 [nucl-ex].
[94] S. Chatrchyan *et al.* [CMS Collaboration], Phys. Rev. Lett. **109** (2012) 222301.
[95] I. Tserruya, arXiv:1311.4456 [nucl-ex].
[96] C. Gale, S. Jeon, B. Schenke, P. Tribedy and R. Venugopalan, Phys. Rev. Lett. **110** (2013) 012302.
[97] V. D. Kekelidze *et al.* [NICA and MPD Collaborations], Phys. Atom. Nucl. **75** (2012) 542.
[98] A. Accardi *et al.*, arXiv:1212.1701 [nucl-ex].
[99] B. Müller, J. Schukraft, and B. Wyslouch, Ann. Rev. Nucl. Part. Sci. **62**, 361 (2012).
[100] B. V. Jacak and B. Müller, Science **337**, 310 (2012).
[101] B. Müller, arXiv:1309.7616 [nucl-th].
[102] J. Schukraft, arXiv:1311.1429 [hep-ex].

Particle Physics Instrumentation

Werner Riegler
CERN, Geneva, Switzerland

Abstract
This report summarizes a series of three lectures aimed at giving an overview of basic particle detection principles, the interaction of particles with matter, the application of these principles in modern detector systems, as well techniques to read out detector signals in high-rate experiments.

1 Introduction

"New directions in science are launched by new tools much more often than by new concepts" is a famous quote from Freeman Dyson's book *Imagined Worlds*. This is certainly true for the field of particle physics, where new tools such as the cloud chamber, bubble chamber, wire chamber, solid-state detectors, accelerators, etc. have allowed physicists to enter into unchartered territory and to discover unexpected phenomena, the understanding of which has provided a deeper insight into the nature of matter. Looking at all Nobel Prize winners connected to the Standard Model of particle physics, one finds many more experimentalists and "instrumentalists" than theoretically orientated physicists, which is a strong indicator of the essence of new tools for advancing our knowledge.

This report will first discuss a few detector systems in order to illustrate the detector needs and specifications of modern particle physics experiments. Then the interaction of particles with matter, which is of course at the heart of particle detection, will be reviewed. Techniques for tracking with gas detectors and solid-state detectors as well as energy measurement with calorimeters are then elaborated. Finally, the tricks on how to process the signals from these detectors in modern high-rate applications will be discussed.

2 Examples of detector systems

The Large Hadron Collider (LHC) experiments ATLAS, CMS, ALICE and LHCb are currently some of the most prominent detectors because of their size, complexity and rate capability. Huge magnet systems, which are used to bend the charged particles in order to measure their momenta, dominate the mechanical structures of these experiments. Proton collision rates of 1 GHz, producing particles and jets of TeV-scale energy, present severe demands in terms of spectrometer and calorimeter size, rate capability and radiation resistance. The fact that only about 100 of the 10^9 events per second can be written to disk necessitates highly complex online event selection, i.e. "triggering'. The basic layout of these collider experiments is quite similar. Close to the interaction point there are several layers of pixel detectors that allow the collision vertices to be distinguished and measured with precision on the tens of micrometres level. This also allows short-lived B and D mesons to be identified by their displaced decay vertices. In order to follow the tracks along their curved path up to the calorimeter, a few metres distant from the collision point, one typically uses silicon strip detectors or gas detectors at larger radii. CMS has an "all-silicon tracker" up to the calorimeter, while the other experiments use also gas detectors like so-called straw tubes or a time projection chamber. The trackers are then followed by the electromagnetic and hadron calorimeter, which measures the energy of electrons, photons and hadrons by completely absorbing them in very large amounts of material. The muons, the only particles able to pass through the calorimeters, are then measured at even larger radii by dedicated muon systems. The sequence of vertex detector, tracker for momentum spectrometry, calorimeter for energy measurement followed again by tracking for muons is the classic basic geometry that underlies most collider and even fixed-target experiments. It allows one to distinguish electrons, photons, hadrons and muons and to measure their momenta and energies.

The ALICE and LHCb experiments use a few additional detector systems that allow different hadrons to be distinguished. By measuring the particle's velocity in addition to the momentum, one can identify the mass and therefore the type of hadron. This velocity can be determined by measuring time of flight, the Cerenkov angle or the particle's energy loss. ALICE uses, in addition, the transition radiation effect to separate electrons from hadrons, and has therefore implemented almost all known tricks for particle identification. Another particle detector using all these well-established techniques is the Alpha Magnetic Spectrometer (AMS) that has recently been installed on the *International Space Station*. It is aimed at measuring the primary cosmic-ray composition and energy distribution.

More "exotic" detector geometries are used for neutrino experiments, which demand huge detector masses in order to make the neutrinos interact. The IceCube experiment at the South Pole uses one cubic kilometre of ice as the neutrino detection medium to look for neutrino point sources in the Universe. Neutrinos passing through the Earth from the Northern Hemisphere interact deep down under the ice and the resulting charged particles are travelling upwards at speeds larger than the speed of light in the ice. They therefore produce Cerenkov radiation, which is detected by a series of more than 5000 photon detectors that are immersed into the ice and look downwards. An example of an accelerator-based neutrino experiment is the CERN Neutrino to Gran Sasso (CNGS) beam. A neutrino beam is sent from CERN over a distance of 732 km to the Gran Sasso laboratory in Italy, where some large neutrino detectors are set up. One of them, the OPERA detector, uses more than 150 000 lead bricks as neutrino target. The bricks are built up from alternating sheets of lead and photographic emulsion, which allows tracking with the micrometre precision necessary to identify the tau leptons that are being produced by interaction of tau neutrinos. This "passive" detector is followed by trigger and tracking devices, which detect secondary particles from the neutrino interactions in the lead bricks and identify the bricks where an interesting event has taken place. To analyse the event, the bricks have then to be removed from the assembly and the photographic emulsion must be developed.

These are only a few examples from a large variety of existing detector systems. It is, however, important to bear in mind that there are only a few basic principles of particle interaction with matter that underly all these different detectors. It is therefore worth going through them in detail.

3 Basics of particle detection

The Standard Model of particle physics counts 17 particles, namely six quarks, six leptons, photon, gluon, W and Z bosons, and the hypothetical Higgs particle. Quarks, however, are not seen as free particles; rather, they combine into baryons and mesons, of which there are hundreds. How can we therefore distinguish all these different particle types in our detectors? The important fact is that, out of the hundreds of known hadrons, only 27 have a lifetime that is long enough such that they can leave a track $> 1\ \mu\mathrm{m}$ in the detector. All the others decay "on the spot" and can only be identified and reconstructed through kinematic relations of their decay products like the "invariant mass". Out of these 27 particles, 13 have lifetimes that make them decay after a distance between a few hundred micrometres and a few millimetres at GeV energies, so they can be identified by their decay vertices, which are only a short distance from the primary collision vertex (secondary vertex tagging). The 14 remaining particles are the only ones that can actually "fly" though the entire detector, and the following eight are by far the most frequent ones: electron, muon, photon, charged pion, charged kaon, neutral kaon, proton and neutron. The principle task of a particle detector is therefore to identify and measure the energies and momenta of these eight particles.

Their differences in mass, charge and type of interaction are the key to their identification, which will be discussed in detail later. The electron leaves a track in the tracking detector and produces a shower in the electromagnetic (EM) calorimeter. The photon does not leave a track but also produces a shower in the EM calorimeter. The charged pion, charged kaon and the proton show up in the tracker but pass through the EM calorimeter and produce hadron showers in the hadron calorimeter. The neutral kaon and the neutron do not show tracks and shower in the hadron calorimeter. The muon is the only particle than

manages to pass through even the hadron calorimeter and is identified by tracking detectors behind the calorimeters. How to distinguish between pion, kaon and proton is typically the task of specific particle identification (PID) detectors.

4 Interaction of particles with matter

The processes leading to signals in particle detectors are now quite well understood and, as a result of available computing power and simulation programs like GEANT or GARFIELD, one can simulate detector responses to the level of a few percent based on fundamental microphysics processes (atomic and nuclear cross-sections). By knowing the basic principles and performing some "back-of-the-envelope calculations", it is possible to estimate detector response to the 20–30% level.

It sounds obvious that any device that is to detect a particle must interact with it in some way. In accelerator experiments, however, there is a way to detect neutrinos even if they do not interact in the detector. Since the total momentum of the colliding particles is known, the sum of all momenta of the produced particles must amount to the same number, owing to momentum conservation. If one uses a hermetic detector, the measurement of missing momentum can therefore be used to detect the momentum vector of the neutrino!

The electromagnetic interaction of charged particles with matter lies at the heart of all particle detection. We can distinguish six types of these interactions: atomic excitation, atomic ionization, bremsstrahlung, multiple scattering, Cerenkov radiation and transition radiation. We will discuss them in more detail in the following.

4.1 Ionization and excitation

A charged particle passing through an atom will interact through the Coulomb force with the atomic electrons and the nucleus. The energy transferred to the electrons is about 4000 times larger compared to the energy transferred to the nucleus because of the much higher mass of the nucleus. We can therefore assume that energy is transferred only to the electrons. In a distant encounter between a passing particle and an electron, the energy transfer will be small – the electron will not be liberated from the atom but will just go to an excited state. In a close encounter the energy transfer can be large enough to exceed the binding energy – the atom is ionized and the electron is liberated. The photons resulting from de-excitation of the atoms and the ionization electrons and ions are used in particle detectors to generate signals that can be read out with appropriate readout electronics.

The faster the particle is passing through the material, the less time there is for the Coulomb force to act, and the energy transfer for the non-relativistic regime therefore decreases with particle velocity v as $1/v^2$. If the particle velocity reaches the speed of light, this decrease should stop and stay at a minimum plateau. After a minimum for Lorentz factors $\gamma = 1/\sqrt{1 - v^2/c^2}$ of ≈ 3, however, the energy loss increases again because the kinematically allowed maximum energy that can be transferred from the incoming particle to the atomic electron is increasing. This rise goes with $\log \gamma$ and is therefore called the relativistic rise. Bethe and Bloch devised a quantum-mechanical calculation of this energy loss in the 1930s. For ultra-relativistic particles, the very strong transverse field will polarize the material and the energy loss will be slightly reduced.

The energy loss is, in addition, independent of the mass of the incoming particle. Dividing the energy loss by the density of the material, it becomes an almost universal curve for all materials. The energy loss of a particle with $\gamma \approx 3$ is around 1–2 $\times \rho[\mathrm{g/cm^3}]$ MeV/cm. Taking iron as an example, the energy for a high-energy particle due to ionization and excitation is about 1 GeV/m. The energy loss is also proportional to the square of the particle charge, so a helium nucleus will deposit four times more energy compared to a proton of the same velocity.

Dividing this energy loss by the ionization energy of the material, we can get a good estimate of the number of electrons and ions that are produced in the material along the track of the passing particle.

Since the energy deposited is a function of the particle's velocity only, we can use it to identify particles: measuring the momentum by the bending in a magnetic field and the velocity from the energy loss, we can determine the mass of the particle in certain momentum regions.

If a particle is stopped in a material, the fact that the energy loss of charged particles increases for smaller velocities results in large energy deposits at the end of the particle track. This is the basis of hadron therapy, where charged particles are used for tumour treatment. These particles deposit a large amount of dose inside the body at the location of the tumour without exposing the overlying tissue to high radiation loads.

This energy loss is, of course, a statistical process, so the actual energy loss will show fluctuations around the average given by the Bethe–Bloch description. This energy-loss distribution was first described by Landau and it shows a quite asymmetric tail towards large values of the energy loss. This large fluctuation of the energy loss is one of the important limiting factors of tracking detector resolution.

4.2 Multiple scattering, bremsstrahlung and pair production

The Coulomb interaction of an incoming particle with the atomic nuclei of the detector material results in deflection of the particle, which is called multiple scattering. A particle entering a piece of material perpendicular to the surface will therefore have a probability of exiting at a different angle, which has a Gaussian distribution with a standard deviation that depends on the particle's properties and the material. This standard deviation is inversely proportional to the particle velocity and the particle momentum, so evidently the effect of multiple scattering and related loss of tracking resolution and therefore momentum resolution is worst for low-energy particles. The standard deviation of the angular deflection is, in addition, proportional to the square root of the material thickness, so clearly one wants to use the thinnest possible tracking devices. The material properties are summarized in the so-called radiation length X_0, and the standard deviation depends on the inverse root of that. Materials with small radiation length are therefore not well suited to the volume of tracking devices. This radiation length X_0 is proportional to $A/\rho Z^2$ where A, ρ and Z are the nuclear number, density and atomic number of the material. Tracking systems therefore favour materials with very low atomic number like beryllium for beampipes, carbon fibre and aluminium for support structures, and thin silicon detectors or gas detectors as tracking elements.

The deflection of the charged particle by the nuclei results in acceleration and therefore emission of electromagnetic radiation. This effect is called "bremsstrahlung" and it plays a key role in calorimetric measurements. The energy loss of a particle due to bremsstrahlung is proportional to the particle energy and inversely proportional to the square of the particle mass. Since electrons and positrons are very light, they are the only particles where energy loss due to bremsstrahlung can dominate over energy loss due to ionization at typical present accelerator energies. The energy of a high-energy electron or positron travelling a distance x in a material decreases as $\exp(-x/X_0)$, where X_0 is again the above-mentioned radiation length. The muon, the next lightest particle, has about 200 times the electron mass, so the energy loss from bremsstrahlung is 40 000 times smaller at a given particle energy. A muon must therefore have an energy of more than 400 GeV in order to have an energy loss from bremsstrahlung that dominates over the ionization loss. This fact can be used to distinguish them from other particles, and it is at the basis of electromagnetic calorimetry through a related effect, the so-called pair production.

A high-energy photon has a certain probability of converting into an electron–positron pair in the vicinity of a nucleus. This effect is closely related to bremsstrahlung. The average distance that a high-energy photon travels in a material before converting into an electron–positron pair is also approximately given by the radiation length X_0. The alternating processes of bremsstrahlung and pair production result in an electromagnetic cascade (shower) of more and more electrons and positrons with increasingly degraded energy until they are stopped in the material by ionization energy loss. We will come back to this in the discussion of calorimetry.

4.3 Cerenkov radiation

Charged particles passing through material at velocities larger than the speed of light in the material produce an electromagnetic shock wave that materializes as electromagnetic radiation in the visible and ultraviolet range, the so-called Cerenkov radiation. With n being the refractive index of the material, the speed of light in the material is c/n, so the fact that a particle does or does not produce Cerenkov radiation can be used to apply a threshold to its velocity. This radiation is emitted at a characteristic angle with respect to particle direction. This Cerenkov angle Θ_c is related to the particle velocity v by $\cos\Theta_c = c/nv$, so by measuring this angle, one can determine the velocity of a charged particle.

4.4 Transition radiation

Transition radiation is emitted when a charged particle crosses the boundary between two materials of different permittivity. The probability of emission is proportional to the Lorentz factor γ of the particle and is only appreciable for ultra-relativistic particles, so it is mainly used to distinguish electrons from other hadrons. As an example a particle with $\gamma = 1000$ has a probability of about 1% to emit a photon on the transition between two materials, so one has to place many layers of material in the form of sheets, foam or fibres in order to produce a measurable amount of radiation. The energy of the emitted photons is in the keV region, so the fact that a charged particle is accompanied by X-rays is used to identify it as an electron or positron.

5 Detector principles

In the previous section we have seen how charged particles leave a trail of excited atoms and electron–ion pairs along their track. Now we can discuss how this is used to detect and measure them. We will first discuss detectors based on atomic excitation, so-called scintillators, where the de-excitation produces photons, which are reflected to appropriate photon detectors. Then we discuss gaseous and solid-state detectors based on ionization, where the electrons and ions (holes) drift in electric fields, which induces signals on metallic readout electrodes connected to readout electronics.

5.1 Detectors based on scintillation

The light resulting from complex de-excitation processes is typically in the ultraviolet to visible range. The three important classes of scintillators are the noble gases, inorganic crystals and polycyclic hydrocarbons (plastics). The noble gases show scintillation even in their liquid phase. An application of this effect is the liquid argon time projection chamber where the instantaneous light resulting from the passage of the particle can be used to mark the start signal for the drift-time measurement. Inorganic crystals show the largest light yield and are therefore used for precision energy measurement in calorimetry applications and also in nuclear medicine. Plastics constitute the most important class of scintillators owing to their cheap industrial production, robustness and mechanical stability. The light yield of scintillators is typically a few percent of the energy loss. In 1 cm of plastic scintillator, a high-energy particle typically loses 1.5 MeV, of which 15 keV goes into visible light, resulting in about 15 000 photons. In addition to the light yield, the decay time, i.e. the de-excitation time, is an important parameter of the scintillator. Many inorganic crystals such as NaI or CsI show very good light yield, but have decay times of tens, even hundreds, of nanoseconds, so they have to be carefully chosen considering the rate requirements of the experiments. Plastic scintillators, on the other hand, are very fast and have decay times on only the nanosecond scale, and they are therefore often used for precision timing and triggering purposes.

The photons produced inside a scintillator are internally reflected to the sides of the material, where so-called “light guides” are attached to guide the photons to appropriate photon detection devices. A very efficient way to extract the light is to use so-called wavelength shifting fibres, which are attached to the side of the scintillator materials. The light entering the fibre from the scintillator is converted into

a longer wavelength there and it can therefore not reflect back into the scintillator. The light stays in the fibre and is internally reflected to the end, where again the photon detector is placed.

The classic device used to convert these photons into electrical signals is the so-called photomultiplier. A photon hits a photocathode, a material with very small work function, and an electron is liberated. This electron is accelerated in a strong electric field to a dynode, which is made from a material with high secondary electron yield. The one electron hitting the surface will therefore create several electrons, which are again guided to the next dynode, and so on, so that out of the single initial electron one ends up with a sizeable signal of, for example, 10^7–10^8 electrons.

In recent years, the use of solid-state photomultipliers, the so-called avalanche photodiodes (APDs), has become very popular, owing to their much lower price and insensitivity to magnetic fields.

5.2 Gaseous detectors

A high-energy particle leaves about 80 electron–ion pairs in 1 cm of argon, which is not enough charge to be detected above the readout electronics noise of typically a few hundred to a few thousand electrons, depending on the detector capacitance and electronics design. A sizeable signal is only seen if a few tens or hundreds of particles cross the gas volume at the same time, and in this operational mode such a gas detector, consisting of two parallel metal electrodes with a potential applied to one of them, is called an "ionization chamber". In order to be sensitive to single particles, a gas detector must have internal electron multiplication. This is accomplished most easily in the wire chamber. Wires of very small diameter, between 10 and 100 μm, are placed between two metallic plates a few millimetres apart. The wires are at a high voltage of a few kilovolts, which results in a very high electric field close to the wire surface. The ionization electrons move towards the thin wires, and, in the strong fields close to the wires, the electrons are accelerated to energies above the ionization energy of the gas, which results in secondary electrons and as a consequence an electron avalanche. Gas gains of 10^4–10^5 are typically used, which makes the wire chambers perfectly sensitive to single tracks. In this basic application, the position of the track is therefore given by the position of the wire that carries a signal, so we have a one-dimensional positioning device.

One has to keep in mind that the signal in the wire is not due to the electrons entering wire; rather, the signal is induced while the electrons are moving towards the wire and the ions are moving away from it. Once all charges arrive at the electrode, the signal is terminated. The signals in detectors based on ionization are therefore *induced* on the readout electrodes by the *movement* of the charges. This means that we find signals not only on electrodes that receive charges but also on other electrodes in the detector. For the wire chamber one can therefore segment the metal plates (cathodes) into strips in order to find the second coordinate of the track along the wire direction. In many applications, one does not even read out the wire signals but instead one segments the cathode planes into square or rectangular pads to get the full two-dimensional information from the cathode pad readout. The position resolution is in this case not limited by the pad size. If one uses pad dimensions of the order of the cathode-to-wire distance, one finds signals on a few neighbouring pads, and, by using centre-of-gravity interpolation, one can determine the track position, which is only 1/10 to 1/100 of the pad size. Position resolution down to 50 μm and rate capabilities of hundreds of kHz of particles per cm^2 per second can be achieved with these devices.

Another way to achieve position resolution that is far smaller than the wire separation is the so-called drift chamber. One determines the time when the particle passes the detector by an external device, which can be a scintillator or the accelerator clock in a collider experiment, and one uses the arrival time of the ionization electrons at the wire as the measure of the distance between the track and the wire. The ATLAS muons system, for instance, uses tubes of 15 mm radius with a central wire, and the measurement of the drift time determines the track position to 80 μm precision.

The choice of the gas for a given gas detector is dominated by the transport properties of electrons

and ions in gases, because these determine the signal and timing characteristics. In order to avoid the ionization electrons getting lost on their way to the readout wires, one can use only gases with very small electronegativity. The main component of detector gases are therefore the noble gases like argon or neon. Other admixtures like hydrocarbons (methane, isobutane) or CO_2 are also needed in order to "tune" the gas transport properties and to ensure operational stability. Since hydrocarbons were shown to cause severe chamber ageing effects at high rates, the LHC detectors use almost exclusively argon, neon and xenon together with CO_2 for all wire chambers.

Typical drift velocities of electrons are in the range of 5–10 cm/μs. The velocity of the ions that are produced in the electron avalanche at the wire and are moving back to the cathodes is about 1000–5000 times smaller than the electron velocity. The movement of these ions produced long signal tails in wire chambers, which have to be properly removed by dedicated filter electronics.

During the past 10–15 years a very large variety of new gas detectors have entered particle physics instrumentation, the so-called micropattern gas detectors like the GEM (gas electron multiplier) or the MICROMEGA (micro mesh gas detector). In these detectors the high fields for electron multiplication are produced by micropattern structures that are realized with photolithographic methods. Their main advantages are rate capabilities far in excess of those achievable in wire chambers, low material budget construction and semi-industrial production possibilities.

5.3 Solid-state detectors

In gaseous detectors, a charged particle liberates electrons from the atoms, which are freely bouncing between the gas atoms. An applied electric field makes the electrons and ions move, which induces signals on the metal readout electrodes. For individual gas atoms, the electron energy levels are discrete.

In solids (crystals), the electron energy levels are in "bands". Inner-shell electrons, in the lower energy bands, are closely bound to the individual atoms and always stay with "their" atoms. However, in a crystal there are energy bands that are still bound states of the crystal, but they belong to the entire crystal. Electrons in these bands and the holes in the lower band can move freely around the crystal, if an electric field is applied. The lowest of these bands is called the "conduction band".

If the conduction band is filled, the crystal is a conductor. If the conduction band is empty and "far away" from the last filled band, the valence band, the crystal is an insulator. If the conduction band is empty but the distance to the valence band is small, the crystal is called a semiconductor.

The energy gap between the valence band and the conduction band is called the band gap E_g. The band gaps of diamond, silicon and germanium are 5.5, 1.12 and 0.66 eV, respectively. If an electron in the valence band gains energy by some process, it can be excited into the conduction band and a hole in the valence band is left behind. Such a process can be the passage of a charged particle, but also thermal excitation with a probability proportional to $\exp(-E_g/kT)$. The number of electrons in the conduction band therefore increases with temperature, i.e. the conductivity of a semiconductor increases with temperature.

It is possible to treat electrons in the conduction band and holes in the valence band similar to free particles, but with an effective mass different from elementary electrons not embedded in the lattice. This mass is furthermore dependent on other parameters such as the direction of movement with respect to the crystal axis. If we want to use a semiconductor as a detector for charged particles, the number of charge carriers in the conduction band due to thermal excitation must be smaller than the number of charge carriers in the conduction band produced by the passage of a charged particle. Diamond can be used for particle detection at room temperature; silicon and germanium must be cooled, or the free charge carriers must be eliminated by other tricks like "doping".

The average energy to produce an electron–hole pair for diamond, silicon and germanium, respectively, is 13, 3.6 and 2.9 eV. Compared to gas detectors, the density of a solid is about a factor of 1000 larger than that of a gas, and the energy to produce an electron–hole pair for silicon, for example, is

a factor 7 smaller than the energy to produce an electron–ion pair in argon. The number of primary charges in a silicon detector is therefore about 10^4 times larger than in a gas and, as a result, solid-state detectors do not need internal amplification. While, in gaseous detectors, the velocities of electrons and ions differ by a factor of 1000, the velocities of electrons and holes in many semiconductor detectors are quite similar, which results in very short signals of a few tens of nanosecond length.

The diamond detector works like a solid-state ionization chamber. One places diamond of a few hundred micrometres thickness between two metal electrodes and applies an electric field. The very large electron and hole mobilities of diamond result in very fast and short signals, so, in addition to tracking application, the diamond detectors are used as precision timing devices.

Silicon is the most widely used semiconductor material for particle detection. A high-energy particle produces around 33 000 electron–hole pairs in 300 μm of silicon. At room temperature there are, however, 1.45×10^{10} electron–hole pairs per cm^3. To apply silicon as a particle detector at room temperature, one therefore has to use the technique of "doping". Doping silicon with arsenic makes it an n-type conductor (more electrons than holes); doping silicon with boron makes it a p-type conductor (more holes that electrons). Putting an n-type and p-type conductor in contact realizes a diode.

At a p–n junction the charges are depleted and a zone free of charge carriers is established. By applying a voltage, the depletion zone can be extended to the entire diode, which results in a highly insulating layer. An ionizing particle produces free charge carriers in the diode, which drift in the electric field and therefore induce an electrical signal on the metal electrodes. As silicon is the most commonly used material in the electronics industry, it has one big advantage with respect to other materials, namely highly developed technology.

Strip detectors are a very common application, where the detector is segmented into strips of a few 50–150 μm pitch and the signals are read out on the ends by wire bonding the strips to the readout electronics. The other coordinate can then be determined, either by another strip detector with perpendicular orientation, or by implementing perpendicular strips on the same wafer. This technology is widely used at the LHC, and the CMS tracker uses 445 m^2 of silicon detectors.

In the very-high-multiplicity region close to the collision point, a geometry of crossed strips results in too many "ghost" tracks, and one has to use detectors with a chessboard geometry, so-called pixel detectors, in this region. The major complication is the fact that each of the chessboard pixels must be connected to a separate readout electronics channel. This is achieved by building the readout electronics wafer in the same geometry as the pixel layout and soldering (bump bonding) each of the pixels to its respective amplifier. Pixel systems in excess of 100 million channels are successfully operating at the LHC.

A clear goal of current solid-state detector development is the possibility of integration of the detection element and the readout electronics into a monolithic device.

6 Calorimetry

The energy measurement of charged particles by completely absorbing ("stopping") them is called calorimetry. Electromagnetic (EM) calorimeters measure the energy of electrons and photons. Hadron calorimeters measure the energy of charged and neutral hadrons.

6.1 Electromagnetic calorimeters

As discussed above, high-energy electrons suffer significant bremsstrahlung owing to their small mass. The interplay of bremsstrahlung and pair production will develop a single electron or photon into a shower of electrons and positrons. The energy of these shower particles decreases exponentially until all of them are stopped due to ionization loss. The total amount of ionization produced by the electrons and positrons is then a measure of the particle energy. The characteristic length scale of this shower process is called the radiation length X_0, and in order to fully absorb a photon or electron one typically uses a

thickness of about 25 X_0. One example of such an EM calorimeter at the LHC is the crystal calorimeter of CMS, which uses PbW_4 crystals. The radiation length X_0 of this crystal is 9 mm, so with a length of 22 cm one can fully absorb the high-energy electron and photon showers. In these crystals the light produced by the shower particles is used as the measure of the energy.

Liquid noble gases are the other prominent materials used for EM calorimetry. In these devices, the total amount of ionization is used as a measure of the energy. The NA48 experiment uses a homogeneous calorimeter of liquid krypton, which has a radiation length of 4.7 cm. Liquid argon has a radiation length of 14 cm, so one would need a depth of 350 cm to fully absorb the EM showers. Since this is not practicable, one interleaves the argon with absorber material of smaller radiation length, such as lead, to allow a more compact design of the calorimeter. Such an alternating assembly of absorber material and active detector material is called a sampling calorimeter. Although the energy resolution of such a device is worse compared to a homogeneous calorimeter, for many applications it is good enough. The ATLAS experiment uses such a liquid argon sampling calorimeter. Other calorimeter types use plastic scintillators interleaved with absorber materials.

The energy resolution of calorimeters improves as $1/\sqrt{E}$ where E is the particle energy. This means that the energy measurement becomes "easier" at high-energy colliders. For homogeneous EM calorimeters, energy resolutions of $\sigma_E/E = 1\%/\sqrt{E\,(\mathrm{GeV})}$ are achieved; typical resolutions of sampling calorimeters are in the range of $\sigma_E/E = (10\text{–}20\%)/\sqrt{E\,(\mathrm{GeV})}$.

6.2 Hadron calorimeters

While only electrons and photons have small enough masses to produce significant EM bremsstrahlung, there is a similar "strong-interaction bremsstrahlung effect" for hadrons. High-energy hadrons radiate pions in the vicinity of a nucleus, and a cascade of these pions develops, which also fully absorbs the incident hadron, and the total ionization loss of this cascade is used to measure the particle energy. The length scale of this shower development is the so-called hadronic interaction length λ, which is significantly larger than the radiation length X_0. For iron the radiation length X_0 is 1.7 cm, whereas the hadronic interaction length λ is 17 cm. Hadron calorimeters are therefore significantly larger and heavier than EM calorimeters. The energy resolution of hadron calorimeters is typically worse than that of EM calorimeters because of the more complex shower processes. About 50% of the energy ends up in pions, 20% ends up in nuclear excitation and 30% goes into slow neutrons, which are usually not detected. A fraction of the produced pions consists of π_0, which instantly decay into two photons, which in turn start an EM cascade. The relative fluctuations of all these processes will result in a larger fluctuation of the calorimeter signal and therefore reduced resolution. Hadron calorimeters are also typically realized as sampling calorimeters with lead or steel plates interleaved with scintillators or liquid noble gases. Energy resolutions of $\sigma_E/E = (50\text{–}100\%)/\sqrt{E\,(\mathrm{GeV})}$ are typical.

7 Particle identification

By measuring the trajectory of a particle in a magnetic field, one measures the particle's momentum, so in order to determine the particle type, i.e. the particle's mass, one needs an additional measurement. Electrons, positrons and photons can be identified by electromagnetic calorimetry, and muons can be identified by the fact that they traverse large amounts of material without being absorbed. To distinguish between protons, kaons and pions is a slightly more subtle affair, and it is typically achieved by measuring the particle's velocity in addition to the momentum.

For kinetic energies that are not too far from the rest mass of the particle, the velocity is not yet too close to the speed of light, such that one can measure the velocity by time of flight. With precision timing detectors like scintillators or resistive plate chambers, time resolutions of less than 100 ps are being achieved. For a time-of-flight distance of 1 m, this allows kaon/pion separation up to 1.5 GeV/c, and proton/pion separation up to about 3 GeV/c.

The energy loss of a particle also measures its velocity, so particle identification up to tens of GeV for pions and protons can be achieved. In gas detectors with pad readout and charge interpolation, the signal pulse height is measured for centre-of-gravity interpolation in view of precision tracking. Since the pulse height is a measure of the energy loss, it can in addition be used for particle identification. Time projection chambers are the best examples of combined tracking and particle identification detectors.

For larger velocities, one can use the measurement of the Cerenkov angle to find the particle velocity. This radiation is emitted at a characteristic angle that is uniquely related to the particle velocity. Using short radiators this angle can be determined simply by measuring the radius of the circle produced by the photons in a plane at a given distance from the radiator. Another technique uses a spherical mirror to project the photons emitted along a longer path onto a plane that also forms a circle. Detectors of this type are called ring imaging Cerenkov detectors (RICH). Since only a "handful" of photons are emitted over typical radiator thicknesses, very efficient photon detectors are the key ingredient to Cerenkov detectors. Using very long gas radiators with very small refractive index, kaon/pion separation up to momenta of 200 GeV/c has been achieved.

8 Signal readout

Many different techniques to make particle tracks visible were developed in the last century. The cloud chamber, the bubble chamber and the photographic emulsion were taking actual pictures of the particle tracks. Nowadays we have highly integrated electronic detectors that allow high particle rates to be processed with high precision. Whereas bubble chambers were almost unbeatable in terms of position resolution (down to a few micrometres) and the ability to investigate very complex decay processes, these detectors were only able to record a few events per second, which is not suitable for modern high-rate experiments. The LHC produces 10^9 proton–proton collisions per second, of which, for example, 100 produce W bosons that decay into leptons, 10 produce a top quark pair and 0.1 produce a hypothetical Higgs particle of 100 GeV. Only around 100 of the 10^9 events per second can be written to tape, which still results in petabytes of data per year to be analysed. The techniques to reduce the rate from 10^9 to 100 Hz by selecting only the "interesting" events is the realm of the so-called trigger and data acquisition. With a bunch crossing time of 25 ns, the particles produced in one collision have not even reached the outer perimeter of the detector when the next collision is already taking place. The synchronization of the data belonging to one single collision is therefore another very challenging task. In order to become familiar with the techniques and vocabulary of trigger and data acquisition, we discuss a few examples.

If, for example, we want to measure temperature, we can use the internal clock of a PC to periodically trigger the measurement. If, on the other hand, we want to measure the energy spectrum of the beta-decay electrons of a radioactive nucleus, we need to use the signal itself to trigger the readout. We can split the detector signal caused by the beta electron and use one path to apply a threshold to the signal, which produces a "logic" pulse that can "trigger" the measurement of the pulse height in the second path. Until this trigger signal is produced, one has to "store" the signal somewhere, which is done in the simplest application by a long cable where the signal can propagate.

If we measure the beta electrons, we cannot distinguish the signals from cosmic particles that are traversing the detector. By building a box around our detector that is made from scintillator, for example, we can determine whether a cosmic particle has entered the detector or whether it was a genuine beta-decay electron. Triggering the readout on the condition of a detector signal in coincidence with the absence of a signal in the scintillator box, we can therefore arrive at a pure beta spectrum sample.

Another example of a simple "trigger" logic is the measurement of the muon lifetime with a stack of three scintillators. Many of the cosmic muons will pass through all three scintillators, but some of them will have lower energy such that they traverse the first one and get stuck in the central one. After a certain time the muon will decay and the decay electron produces a signal in the central and the bottom scintillators. By starting a clock with a signal condition of 1 AND 2 ANDNOT 3 and stopping the clock

with NOT 1 AND 2 AND 3, one can measure the lifetime of the muons.

At the LHC experiment some typical trigger signals are high-energy events transverse to the proton beam direction, which signify interesting high-energy parton collisions. High-energy clusters in the calorimeters or high-energy muons are therefore typical trigger signals, which start the detector readout and ship the data to dedicated processing units for further selection refinement.

In order to cope with high rates, one has to find appropriate ways to deal with the "processing" time, i.e. the time while the electronics is busy with reading out the data. This we discuss in the following. First we assume a temperature sensor connected to a PC. The PC has an internal clock, which can be used to periodically trigger the temperature measurement and write the values to disk. The measurement and data storage will take a certain time τ, so this "deadtime" limits the maximum acquisition rate. For a deadtime $\tau = 1$ ms, we have a maximum acquisition rate of $f = 1/\tau = 1$ kHz.

For the example of the beta spectrum measurement, we are faced with the fact that the events are completely random and it can happen that another beta decay takes place while the acquisition of the previous one is still ongoing. In order to avoid triggering the readout while the acquisition of the previous event is still ongoing, one has to introduce a so-called "busy logic", which blocks the trigger while the readout is ongoing. Because the time between events typically follows an exponential distribution, there will always be events lost even if the acquisition time is smaller than the average rate of events. In order to collect 99% of the events, one has to overdesign the readout system with a deadtime of only 10% of the average time between events. To avoid this problem, one uses a so-called FIFO (first-in first-out) buffer in the data stream. This buffer receives as input the randomly arriving data and stores them in a queue. The readout of the buffer happens at constant rate, so by properly choosing the depth of the buffer and the readout rate, it is possible to accept all data without loss, even for readout rates close to the average event rate. This transformation from random input to clocked output is call "de-randomization".

In order to avoid "storing" the signals in long cables, one can also replace them by FIFOs. At colliders, where the bunch crossing comes in regular intervals, the data are stored in so-called front-end pipelines, which sample the signals at the bunch crossing rate and store them until a trigger decision arrives.

The event selection is typically performed at several levels of increasing refinement. The fast trigger decisions in the LHC experiments are performed by specialized hardware on or close to the detector. After a coarse events selection, the rates are typically low enough to allow a more refined selection using dedicated computer farms that do more sophisticated analysis of the events. The increasing computing power, however, drives the concepts of trigger and data acquisition into quite new directions. The concepts for some future high-energy experiments foresee so-called "asynchronous" data-driven readout concepts, where the signal of each detector element receives a time stamp and is then shipped to a computer farm where the event synchronization and events selection is carried out purely by software algorithms.

Probability and Statistics for Particle Physicists

José Ocariz
Université Paris-Diderot and Laboratoire de physique nucléaire et des hautes énergies LPNHE
CERN-IN2P3, Paris, France

Abstract
A pedagogical selection of topics in probability and statistics is presented. Choice and emphasis are driven by the author's personal experience, predominantly in the context of physics analyses using experimental data from high-energy physics detectors.

1 Introduction

These notes are based on a series of three lectures on probability and statistics, given at the AEPSHEP 2012 physics school. While most of the attendance was composed of PhD students, it also included Master students and young post-docs; in consequence, a variable level of familiarity with the topics discussed was implicit. For consistency, the scope of the lectures spanned from very general concepts up to more advanced, recent developments. The first lecture reviewed basic concepts in probability and statistics; the second lecture focussed on maximum likelihood and multivariate techniques for statistical analysis of experimental data; the third and last lecture covered topics on hypothesis-testing and interval estimation. Whenever possible, the notation aligns with common usage in experimental high-energy physics (HEP), and the discussion is illustrated with examples related to recent physics results, mostly from the B-factories and the LHC experiments.

2 Basic concepts in probability and statistics

Mathematical probability is an abstract axiomatic concept, and probability theory is the conceptual framework to assess the knowledge of random processes. A detailed discussion of its development and formalism lies outside the scope of these notes. Other than standard classic books, like [1], there are excellent references available, often written by (high-energy) physicists, and well-suited for the needs of physicists. A non-comprehensive list includes [2–5], and can guide the reader into more advanced topics. The sections on statistics and probability in the PDG [6] are also a useful reference; often also, the large experimental collaborations have (internal) forums and working groups, with many useful links and references.

2.1 Random processes

For a process to be called random, two main conditions are required: its outcome cannot be predicted with complete certainty, and if the process is repeated under the very same conditions, the new resulting outcomes can be different each time. In the context of experimental particle physics, such an outcome could be "a collision", or "a decay". In practice, the sources of uncertainty leading to random processes can be

- due to reducible measurement errors, i.e. practical limitations that can in principle be overcome by means of higher-performance instruments or improved control of experimental conditions;
- due to quasi-irreducible random measurement errors, i.e. thermal effects;
- fundamental, if the underlying physics is intrinsically uncertain, i.e. quantum mechanics is not a deterministic theory.

Obviously in particle physics, all three kinds of uncertainties are at play. A key feature of collider physics is that events resulting from particle collisions are independent of each other, and provide a quasi-perfect laboratory of quantum-mechanical probability processes. Similarly, unstable particles produced in HEP experiments obey quantum-mechanical decay probabilities.

2.2 Mathematical probability

Let Ω be the total universe of possible outcomes of a random process, and let $X, Y \ldots$ be elements (or realizations) of Ω; a set of such realizations is called a sample. A probability function $\mathcal{P}$ is defined as a map onto the real numbers:

$$\begin{aligned} \mathcal{P} : \{\Omega\} &\rightarrow [0:1] \ , \\ X &\rightarrow \mathcal{P}(X) \ . \end{aligned} \tag{1}$$

This mapping must satisfy the following axioms:

$$\begin{aligned} \mathcal{P}(\Omega) &= 1 \ , \\ \text{if } X \cap Y &= \oslash \ , \text{ then } \mathcal{P}(X \cup Y) = \mathcal{P}(X) + \mathcal{P}(Y) \ , \end{aligned} \tag{2}$$

from which various useful properties can be easily derived, i.e.

$$\begin{aligned} \mathcal{P}(\overline{X}) &= 1 - \mathcal{P}(X) \ , \\ \mathcal{P}(X \cup \overline{X}) &= 1 \ , \\ \mathcal{P}(\oslash) &= 1 - \mathcal{P}(\Omega) = 0 \ , \\ \mathcal{P}(X \cup Y) &= \mathcal{P}(X) + \mathcal{P}(Y) - \mathcal{P}(X \cap Y) \ , \end{aligned} \tag{3}$$

(where $\overline{X}$ is the complement of X). Two elements X and Y are said to the independent (that is, their realizations are not linked in any way) if

$$\mathcal{P}(X \cap Y) = \mathcal{P}(X)\mathcal{P}(Y). \tag{4}$$

2.2.1 Conditional probability and Bayes' theorem

Conditional probability $\mathcal{P}(X \mid Y)$ is defined as the probability of X, given Y. The simplest example of conditional probability is for independent outcomes: from the definition of independence in Eq. (4), it follows that if X and Y are actually independent, the condition

$$\mathcal{P}(X \mid Y) = \mathcal{P}(X) \tag{5}$$

is satisfied. The general case is given by Bayes' theorem: in view of the relation $\mathcal{P}(X \cap Y) = \mathcal{P}(Y \cap X)$, it follows that

$$\mathcal{P}(X \mid Y) = \frac{\mathcal{P}(Y \mid X)\mathcal{P}(X)}{\mathcal{P}(Y)} \ . \tag{6}$$

An useful corollary follows as consequence of Bayes' theorem: if Ω can be divided into a number of disjoint subsets X_i (this division process is called a partition), then

$$\mathcal{P}(X \mid Y) = \frac{\mathcal{P}(Y \mid X)\mathcal{P}(X)}{\sum_i \mathcal{P}(Y \mid X_i)\mathcal{P}(X_i)} \ . \tag{7}$$

2.2.2 The probability density function

In the context of these lectures, the relevant scenario is when the outcome of a random process can be stated in numerical form (i.e. it corresponds to a measurement): then to each element X (which for HEP-oriented notation purposes, it is preferable to design as an event) corresponds a variable x (that can be real or integer). For continuous x, its probability density function (PDF) $P(x)$ is defined as

$$\mathcal{P}(X \text{ found in } [x, x+dx]) \;=\; P(x)dx \;, \tag{8}$$

where $P(x)$ is positive-defined for all values of x, and satisfies the normalization condition

$$\int_{-\infty}^{+\infty} dx' P(x') \;=\; 1 \;. \tag{9}$$

For a discrete x_i, the above definition can be adapted in a straightworward way:

$$\begin{aligned} \mathcal{P}(X \text{ found in } x_i) \;&=\; p_i \;, \\ \text{with } \sum_j p_j = 1 \text{ and } p_k \geq 0 \;&\forall k \;. \end{aligned} \tag{10}$$

Finite probabilities are obtained by integration over a non-infinitesimal range. It is sometimes convenient to refer to the cumulative density function (CDF) :

$$C(x) \;=\; \int_{-\infty}^{x} dx' P(x') \;, \tag{11}$$

so that finite probabilities can be obtained by evaluating the CDF on the boundaries of the range of interest :

$$\mathcal{P}(a < X < b) \;=\; C(b) - C(a) \;=\; \int_a^b dx' P(x') \;. \tag{12}$$

Other than the conditions of normalization Eq. (9) and positive-defined (or more precisely, to have a compact support, which implies that the PDF must become vanishingly small outside of some finite boundary) the PDFs can be arbitrary otherwise, and exhibit one or several local maxima or local minima. In contrast, the CDF is a monotonically increasing function of x, as shown on Figure 1, where a generic PDF and its corresponding CDF are represented.

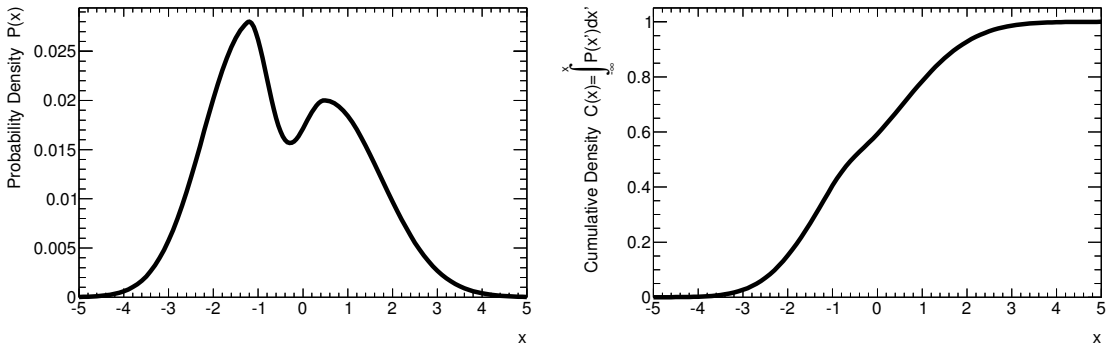


Fig. 1: Left: a probability density function (PDF) for a variable x; the PDF is assumed to have negligible values outside of the plotted range. Right: the corresponding cumulative density function (CDF), plotted in the same range.

2.2.3 Multidimensional PDFs

When more than on random number is produced as outcome in a same event, it is convenient to introduce a n-dimensional set of random elements $\vec{X} = \{X_1, X_2, \ldots, X_n\}$, together with its corresponding set of random variables $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ and their multidimensional PDF:

$$P(\vec{x})d\vec{x} \;=\; P(x_1, x_2, \ldots, x_n)dx_1 dx_2 \ldots dx_n \; . \tag{13}$$

Lower-dimensional PDFs can be derived from Eq. (13); for instance, when one specific variable $x = x_j$ (for fixed j, with $1 \leq j \leq n$) is of particlar relevance, its one-dimensional marginal probability density $P_X(x)$ is extracted by integrating $P(\vec{x})$ over the remaining $n-1$ dimensions (excluding the j-th):

$$P_X(x)dx = dx \int_{-\infty}^{+\infty} dx_1 \ldots \int_{-\infty}^{+\infty} dx_{j-1} \int_{-\infty}^{+\infty} dx_{j+1} \ldots \int_{-\infty}^{+\infty} dy_{n-1} \; . \tag{14}$$

Without loss of generality, the discussion can be restricted to the two-dimensional case, with random elements X and Y and random variables $\vec{X} = \{x, y\}$. The finite probability in a rectangular two-dimensional range is

$$\mathcal{P}(a < X < b \text{ and } c < Y < d) \;=\; \int_a^b dx \int_c^d dy P(x, y) \; . \tag{15}$$

For a fixed value of Y, the conditional density function for X is given by

$$P(x \mid y) = \frac{P(x, y)}{\int dy P(x, y)} = \frac{P(x, y)}{P_Y(y)} \; . \tag{16}$$

As already mentioned, the relation $P(x, y) = P_X(x) \cdot P_Y(y)$ holds only if X and Y are independent; for instance, the two-dimensional density function in Figure 2 is an example of non-independent variables, for which $P(x, y) \neq P_X(x) \cdot P_Y(y)$.

3 Parametric PDFs and parameter estimation

The description of a random process via density functions is called a model. Loosely speaking, a parametric model assumes that its PDFs can be completely described using a finite number of parameters [1]. A straightforward implementation of a parametric PDF is when its parameters are analytical arguments of the density function; the notation $P\,(x, y, \ldots\,; \theta_1, \theta_2, \ldots\,)$ indicates the functional dependence of the PDF (also called its shape) in terms of variables $x_1, y_2, \ldots$ and parameters $\theta_1, \theta_2, \ldots$.

3.1 Expectation values

Consider a random variable X with PDF $P(x)$. For a generic function $f(x)$, its expectation value $E[f]$ is the PDF-weighted average over the x range :

$$E[f] \;=\; \int dx P(x) f(x) \; . \tag{17}$$

Being often used, some common expectation values have their own names. For one-dimensional PDFs, the mean value and variance are defined as

$$\text{Mean value} \quad : \; \mu \;=\; E[x] \;=\; \int dx P(x) x \; , \tag{18}$$

$$\text{Variance} \quad : \; \sigma^2 \;=\; V[x] \;=\; E[x^2] - \mu^2 \;=\; E[(x-\mu)^2] \; ; \tag{19}$$

[1] This requirement needs not to be satisfied; PDFs can also be non-parametric (which is equivalent to assume that an infinite number of parameters is needed to describe them), or they can be a mixture of both types.

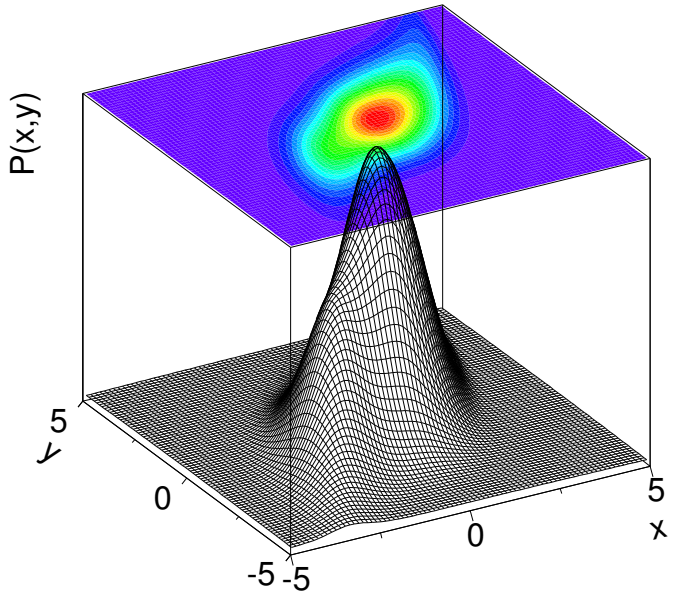


Fig. 2: A two-dimensional probability density function (PDF) for non-independent variables x and y: the pattern implies that in general, the probability densities are larger when x and y are both large or both small (i.e. they are positively correlated), and thus the variables are not independent. The PDF is assumed to have negligible values outside of the plotted two-dimensional range.

for multidimensional PDFs, the covariance $C_{ij} = C(x_i, x_j)$ and the dimensionless linear correlation coefficient ρ_{ij} are defined as:

$$\text{Covariance} \;:\; C_{ij} \;=\; E[x_i x_j] - \mu_i \mu_j \;=\; E[(x_i - \mu_i)(x_j - \mu_j)] \;, \tag{20}$$

$$\text{Linear correlation} \;:\; \rho_{ij} \;=\; \frac{C_{ij}}{\sigma_i \sigma_j} \;. \tag{21}$$

By construction, linear correlation coefficients have values in the $-1 \leq \rho_{ij} \leq 1$ interval. The sign of the ρ coefficient indicates the dominant trend in the $(x_j; x_j)$ pattern: for positive correlation, the probability density is larger when x_i and x_j are both small or large, while a negative correlation indicates that large values of x_i are preferred when small values of x_j are realized (and viceversa). When the random variables X_i and X_j are independent, that is $P(x_i, x_j) = P_{X_i}(x_i) P_{X_j}(x_j)$, one has

$$E[x_i x_j] \;=\; \int\int dx_i dx_j P(x_i, x_j) x_i x_j \;=\; \mu_i \mu_j \;, \tag{22}$$

and thus $\rho_{ij} = 0$: independent variables have a zero linear correlation coefficient. Note that the converse needs not be true: non-linear correlation patterns among non-independent variables may "conspire" and yield null values of the linear correlation coefficient, for instance if negative and positive correlation patterns in different regions of the (x_i, x_j) plane cancel out. Figure 3 shows examples of two-dimensional samples, illustrating a few representative correlation patterns among their variables.

3.2 Shape characterisation

In practice, the true probability density function may not be known, and the accessible information can only be extracted from a finite-size sample (say consisting of N events), which is assumed to have been

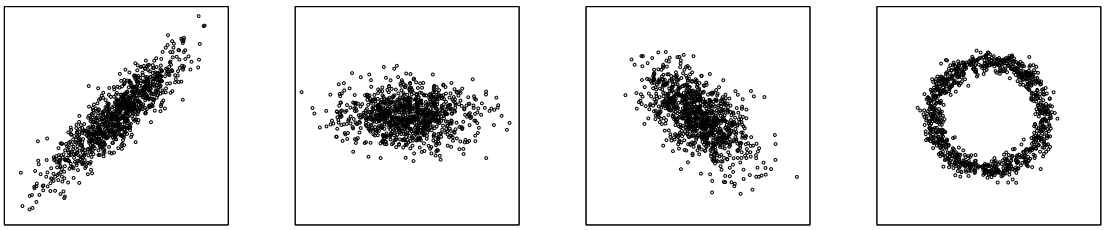


Fig. 3: Examples of two-dimensional probability density functions, illustrating four typical correlation patterns. From left ro right, the figures show two variables which exhibit: a large, positive linear correlation (with $\rho = +0.9$); no correlation, with a zero linear correlation coefficient; a slightly milder, negative correlation (with $\rho = -0.5$); and a more complex correlation pattern, with variables very strongly correlated, but in such a way that the linear correlation coefficient is zero.

originated from an unknown PDF. Assuming that this underlying PDF is parametric, a procedure to estimate its functional dependence and the values of its parameters is called a characterisation of its shape. Now, only a finite number of expectation values can be estimated from a finite-size sample. Therefore, when choosing the set of parameters to be estimated, each should provide information as useful and complementary as possible; such procedure, despite being intrinsically incomplete, can nevertheless prove quite powerful.

The procedure of shape characterisation is first illustrated with the one-dimensional case of a single random variable x. Consider the empirical average $\overline{x}$ (also called sample mean)

$$\overline{x} \;=\; \frac{1}{N}\sum_{i=1}^{N} x_i \;. \tag{23}$$

As shown later in Sec. 3.3.1, $\overline{x}$ is a good estimator of the mean value μ of the underlying distribution $P(x)$. In the same spirit, the quadratic sample mean (or root-mean-square RMS),

$$\mathrm{RMS} \;=\; \sqrt{\overline{x^2} - (\overline{x})^2}\;, \tag{24}$$

is a reasonable estimator of its variance σ^2 (an improved estimator of variance can be easily derived from the RMS, as discussed later in Sec. 3.3.1). Intuitively speaking, these two estimators together provide complementary information on the "location" and "spread" of the region with highest event density in x.

The previous approach can be discussed in a more systematic way, by means of the characteristic function, which is a transformation from the x-dependence of the PDF $P(x)$ onto a k-dependence of $C[k]$, defined as

$$C[k] \;=\; E\left[e^{ik\frac{x-\mu}{\sigma}}\right] \;=\; \sum_{j} \frac{(ik)^j}{j!}\mu_j \;. \tag{25}$$

The coefficients μ of the expansion are called reduced moments; by construction, the first moments are $\mu_1 = 0$ and $\mu_2 = 1$; these values indicate that in terms of the rescaled variable $x' = (x-\mu)/\sigma$, the PDF has been shifted to have zero mean and scaled to have unity variance.

In principle, the larger the number of momenta μ_j that are estimated, the more detailed is the characterisation of the PDF shape. Among higher-order moments, the third and fourth have specific names, and their values can be interpreted, in terms of shape, in a relatively straightforward manner. The third moment is called skewness: a symmetric distribution has zero skewness, and a negative (positive) value

indicates a larger spread to the left (right) of its median. The fourth moment is called kurtosis; it is a positive-defined quantity, that (roughly speaking) can be related to the "peakedness" of a distribution: a large value indicates a sharp peak and long-range tails (such a distribution is sometimes called leptokurtic), while a smaller value reflects a broad central peak with short-range tails (a so-called platykurtic distribution). Figure 4 shows a few distributions, chosen to illustrate the relation of momenta and shapes.

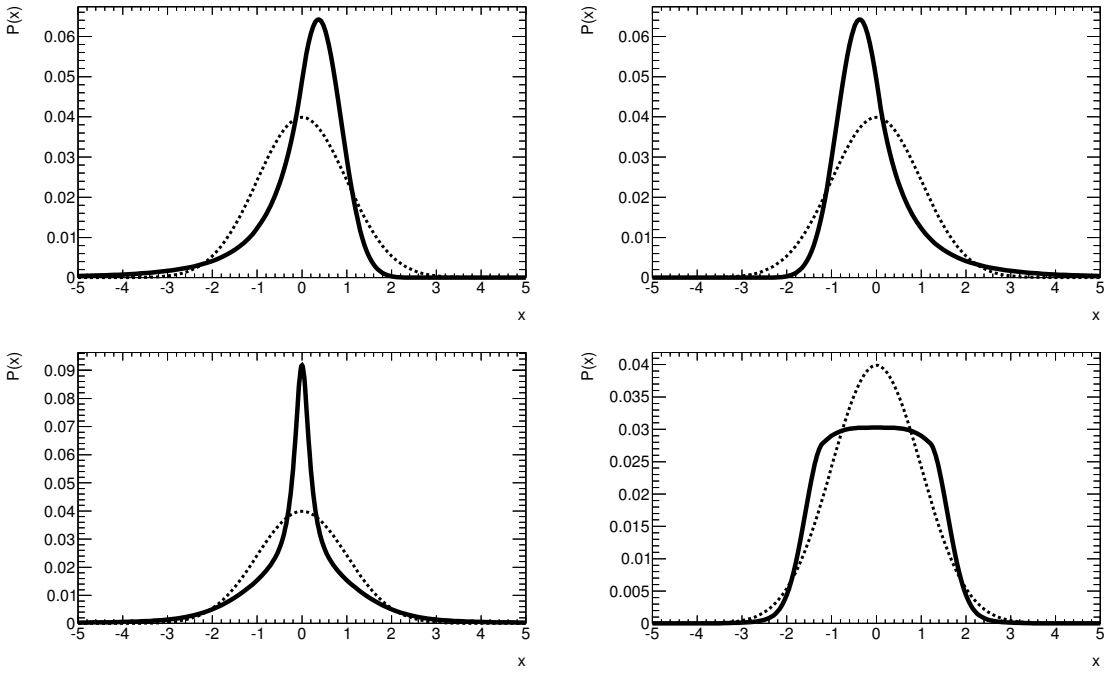


Fig. 4: Examples of probability density functions, illustrating the role of momenta in the characterization of PDF shapes. For all PDFs plotted, their mean values are 0 and their variances are 1. Top plots : the PDF on the left (resp. right) exhibits an asymmetric, non-gaussian tail to the left (resp. right) of its peak, and thus has a positive (resp. negative) skewness. Bottm plots: the PDF on the left (resp. right) shows a narrow peak and long-range tails (resp. broad core and short-range tails), and has a large (resp. small) kurtosis. In each plot, a Gaussian PDF (dashed line) with mean value 0 and variance 1, is also overlaid.

3.3 Parameter estimation

The characterization of the shape of a PDF through a sequential estimation of shape parameters discussed in Sec. 3.2, aimed at a qualitative introduction to the concept of parameter estimation (also called point estimation in the litterature). A more general approach is now discussed in this paragraph.

Consider a n-dimensional, k-parametric PDF,

$$P(\ x_1, x_2, \ldots, x_n\ ;\ \theta_1, \theta_2, \ldots, \theta_k\ )\ , \tag{26}$$

for which the values $\theta_1, \ldots, \theta_k$ are to be estimated on a finite-sized sample, by means of a set of estimators denoted $\hat{\theta_1}, \ldots, \hat{\theta_k}$. The estimators themselves are random variables, with their own mean values and variances: their values differ when estimated on different samples. The estimators should satisfy two key properties: to be consistent and unbaised. Consistency ensures that, in the infinite-sized sample limit, the estimator converges to the true parameter value; absence of bias ensures that the expectation value of the estimator is, for all sample sizes, the true parameter value. A biased, but consistent estimator (also

called asymptotically unbiased) is such that the bias decreases with increasing sample size. Additional criteria can be used to characterize the quality and performance of estimators; two often mentioned are

- efficiency: an estimator with small variance is said to be more efficient than one with larger variance;
- robustness: this criterion characterizes the sensitivity of an estimator to uncertainties in the underlying PDF. For example, the mean value is robust against uncertainties on even-order moments, but is less robust with respect to changes on odd-order ones.

Note that these criteria may sometimes be mutually conflicting; for practical reasons, it may be preferable to use an efficient, biased estimator to an unbiased, but poorly convergent one.

3.3.1 The classical examples: mean value and variance

The convergence and bias requirements can be suitably illustrated with two classical, useful examples, often encountered an many practical situations: the mean value and the variance.

The empirical average $\overline{x}$ is a convergent, unbiased estimation of the mean value μ of its underlying distribution: $\hat{\mu} = \overline{x}$. This statement can easily be demostrated, by evaluating the expectation value and variance of $\overline{x}$:

$$E\left[\overline{x}\right] \;=\; \frac{1}{N}\sum_{i=1}^{N} E\left[x\right] \;=\; \mu \;, \tag{27}$$

$$V\left[\overline{x}\right] \;=\; E\left[(\overline{x}-\mu)^2\right] \;=\; \frac{\sigma^2}{N} \;. \tag{28}$$

In contrast, the empirical RMS of a sample is a biased, asymptotically unbiased estimator of the variance σ^2: this can be demonstrated by first rewriting its square (also called sample variance) in terms of the mean value:

$$\mathrm{RMS}^2 \;=\; \frac{1}{N}\sum_{i=1}^{N}\left(x_i-\overline{x}\right)^2 \;=\; \frac{1}{N}\sum_{i=1}^{N}\left(x_i-\mu\right)^2 - \left(\overline{x}-\mu\right)^2 \;, \tag{29}$$

and so its expectation value is

$$E\left[\mathrm{RMS}^2\right] \;=\; \sigma^2 - V\left[\overline{x}\right] \;=\; \frac{N-1}{N}\sigma^2 \;, \tag{30}$$

which, while properly converging to the true variance σ^2 in the $N \to \infty$ limit, systematically underestimates its value for a finite-sized sample. One can instead define a modified estimator

$$\frac{N}{N-1}\mathrm{RMS}^2 \;=\; \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\overline{x}\right)^2 \;, \tag{31}$$

that ensures, for finite-sized samples, an unbiased estimation of the variance.

In summary, consistent and unbiased estimators for the mean value μ and variance σ^2 of an unknown underlying PDF can be extracted from a finite-sized sample realized out of this PDF:

$$\hat{\mu} \;=\; \frac{1}{N}\sum_{i=1}^{N} x_i \;, \tag{32}$$

$$\hat{\sigma}^2 \;=\; \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\hat{\mu}\right)^2 \;. \tag{33}$$

As a side note, the $1/N$ and $1/(N-1)$ factors for $\hat{\mu}$ in Eq. (32) and for $\hat{\sigma}^2$ in Eq. (33) can be intuitively understood as follows: while the empirical average be estimated even on the smallest sample consisting of a single event, at least two events are needed to estimate their empirical dispersion.

3.3.2 Covariance, correlations, propagation of uncertainties

These two classical examples discussed in 3.3.1 refer to a single random variable. In presence of several random variables, expressed as a random n-dimensional vector $\vec{x} = \{x_1, \ldots, x_n\}$, the discussion leads to the definition of the empirical covariance matrix, whose elements $\hat{C}_{ab}$ can be estimated on a sample of N events as

$$\hat{C}_{ab} \;=\; \frac{1}{N-1} \sum_{i=1}^{N} \left(x_{a,i} - \hat{\mu_a}\right) \left(x_{b,i} - \hat{\mu_b}\right) \;. \tag{34}$$

(the a, b indices run over random variables, $1 \leq a, b \leq n$). Assuming the covariance is known (i.e. by means of the estimator above, or from first principles), the variance of an arbitrary function of these random variables $f(\vec{x})$ can be evaluated from a Taylor-expansion around the mean values $\hat{\vec{\mu}}$ as

$$f\left(\vec{x}\right) \;=\; f\left(\hat{\vec{\mu}}\right) + \sum_{a=1}^{n} \left.\frac{df}{dx_a}\right|_{\vec{x}=\hat{\vec{\mu}}} \left(x_a - \hat{\mu}_a\right) \;, \tag{35}$$

which leads to $E\left[f(\vec{x})\right] \;\simeq\; f(\hat{\vec{\mu}})$; similarly,

$$E\left[f^2(\vec{x})\right] \;\simeq\; f^2(\hat{\vec{\mu}}) + \sum_{a,b=1}^{n} \left.\frac{df}{dx_a}\frac{df}{dx_b}\right|_{\vec{x}=\hat{\vec{\mu}}} \hat{C}_{ab} \;, \tag{36}$$

and thus the variance of f can be estimated as

$$\hat{\sigma}_f^2 \;\simeq\; \sum_{a,b=1}^{n} \left.\frac{df}{dx_a}\frac{df}{dx_b}\right|_{\vec{x}=\hat{\vec{\mu}}} \hat{C}_{ab} \;. \tag{37}$$

This expression in Eq. (37), called the error propagation formula, allows to estimate the variance of a generic function $f(\vec{x})$ from the estimators of mean values and covariances.

A few particular examples of error propagation deserve being mentioning explicitly:

– If all random variables $\{x_a\}$ are uncorrelated, the covariance matrix is diagonal, $C_{ab} = \sigma_a^2 \delta_{ab}$ and the covariance of $f(\vec{x})$ reduces to

$$\hat{\sigma}_f^2 \;\simeq\; \sum_{a=1}^{n} \left.\left(\frac{df}{dx_a}\right)^2\right|_{\vec{x}=\hat{\vec{\mu}}} \hat{\sigma}_a^2 \;. \tag{38}$$

– For the sum of two random variables $S = x_1 + x_2$, the variance is

$$\sigma_S^2 \;=\; \sigma_1^2 + \sigma_2^2 + 2C_{12} \;=\; \sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2\rho_{12}, \tag{39}$$

and the corresponding generalization to more than two variables is straightforward:

$$\sigma_S^2 \;=\; \sum_{a,b} \sigma_a \sigma_b \rho_{ab} \;. \tag{40}$$

In absence of correlations, one says that absolute errors are added in quadrature: hence the expression in Eq. (39) is often written as $\sigma_S \;=\; \sigma_1 \oplus \sigma_2$.

– For the product of two random variables $P = x_1 x_2$, the variance is

$$\left(\frac{\sigma_P}{P}\right)^2 \;=\; \left(\frac{\sigma_1}{x_1}\right)^2 + \left(\frac{\sigma_2}{x_2}\right)^2 + 2\frac{\sigma_1}{x_1}\frac{\sigma_2}{x_2}\rho_{12} \;, \tag{41}$$

with its generalization to more than two variables:

$$\left(\frac{\sigma_P}{P}\right)^2 \;=\; \sum_{a,b} \frac{\sigma_a}{x_a}\frac{\sigma_b}{x_b}\rho_{ab}\,. \tag{42}$$

In absence of correlations, one says that relative errors are added in quadrature, and Eq. (41) $\sigma_P/P = \sigma_1/x_1 \oplus \sigma_2/x_2$.

– For a generic power law function, $Z = x_1^{n_1} x_2^{n_2} \ldots$, if all variables are uncorrelated, the variance is

$$\frac{\sigma_Z}{Z} \;=\; n_1\frac{\sigma_1}{x_1} \oplus n_2\frac{\sigma_2}{x_2} \oplus \ldots \tag{43}$$

4 A survey of selected distributions

In this Section, a brief description of distributions, often encountered in practical applications, is presented. The rationale leading to this choice of PDFs is driven either by their specific mathematical properties, and/or in view of their common usage in the modellingi of important physical processes; such features are correspondingly emphasized in the discussion.

4.1 Two examples of discrete distributions: Binomial and Poisson

4.1.1 The binomial distribution

Consider a scenario with two possible outcomes: "success" or "failure", with a fixed probability p of "success" being realized (this is also called a Bernouilli trial). If n trials are performed, $0 \leq k \leq n$ may actually result in "success"; it is assumed that the sequence of trials is irrelevant, and only the number of "success" k is considered of interest. The integer number k follows the so-called Binomial distribution $P(k; n, p)$:

$$P_{\text{binomial}}\,(k; n, p) \;=\; \frac{n!}{k!(n-k)!} p^k\,(1-p)^{n-k}\;, \tag{44}$$

where k is the random variable, while n and p are parameters. The mean value and variance are

$$\begin{aligned} E\,[k] \;&=\; \sum_{k=1}^{n} kP\,(k; n, p) \;=\; np\,, \\ V\,[k] \;&=\; np(1-p)\,. \end{aligned} \tag{45}$$

4.1.2 The Poisson distribution

In the $n \to \infty$, $p \to 0$ limit (with $\lambda = np$ finite and non-zero) for the Binomial distribution, the random variable k follows the Poisson distribution $P(k; \lambda)$,

$$P_{\text{Poisson}}\,(k; \lambda) \;=\; \frac{\lambda^k e^{-\lambda}}{k!}\;, \tag{46}$$

for which λ is the unique parameter. For Poisson, the mean value and variance are the same:

$$E\,[k] \;=\; V\,[k] \;=\; \lambda\,. \tag{47}$$

The Poisson distribution, sometimes called law of rare events (in view of the $p \to 0$ limit), is a useful model for describing event-counting rates. Examples of a Binomial and a Poisson distribution, for $n = 10$, $p = 0.6$ and for $\lambda = 6.5$, respectively, are shown on Figure 5.

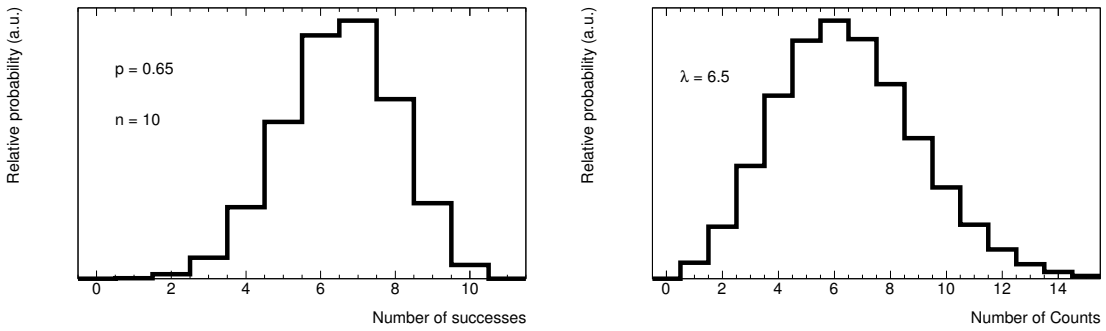


Fig. 5: Examples of a Binomial (left) and a Poisson (right) distributions, with parameters $n = 10$, $p = 0.6$ and $\lambda = 6.5$, respectively.

4.2 Common examples of real-valued distributions

The first two continuous random variables discussed in this paragraph are the uniform and the exponential distributions, illustrated in Figure 6.

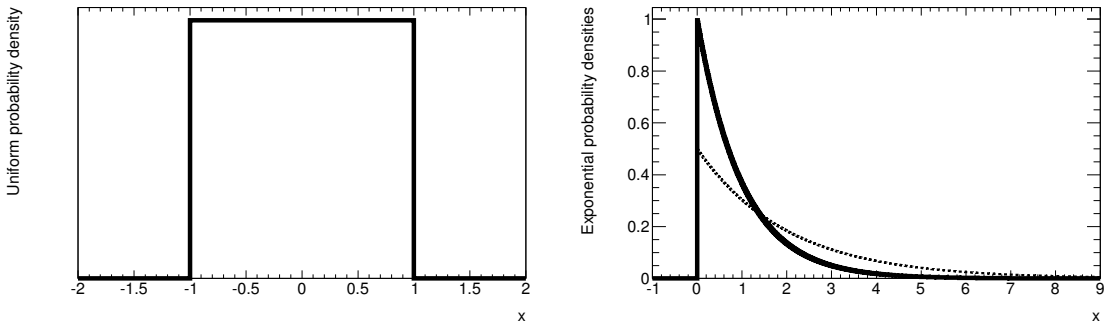


Fig. 6: Examples of a uniform (left) and an exponentially-decreasing (right) distribution. For the uniform distribution, values used for the boundary parameters are $a = -1$, $b = 1$; for the exponential, two values $\xi = 1$ (solid line) and $\xi = 2$ (dashed line) are used.

4.2.1 The uniform distribution

Consider a continuous random variable x, with a probability density $P(x; a, b)$ that is non-zero only inside a finite interval $[a, b]$:

$$P_{\text{uniform}}(x; a, b) = \begin{cases} \frac{1}{b-a} & , \quad a \leq x \leq b \,, \\ 0 & , \quad \text{otherwise} \,. \end{cases} \tag{48}$$

For this uniform distribution, the mean value and variance are

$$E[x] = \frac{a+b}{2} \,, \tag{49}$$

$$V[x] = \frac{(b-a)^2}{12} \,. \tag{50}$$

While not being the most efficient one, a straightforward Monte-Carlo generation approach would be based on a uniform distribution in the $[0, p]$ range, and use randomly generated values of x in the $[0, 1]$ range as implementation of an accept-reject algorithm with success probability p.

4.2.2 The exponential distribution

Consider a continuous variable x, with a probability density $P(x;\xi)$ given by

$$P_{\text{exponential}}\,(x;\xi) \;=\; \begin{cases} \frac{1}{\xi}e^{-x/\xi} & , \quad x \geq 0\,, \\ 0 & , \quad \text{otherwise}\,, \end{cases} \tag{51}$$

whose mean value and variance are

$$E\,[x] \;=\; \xi\,, \tag{52}$$
$$V\,[x] \;=\; \xi^2\,. \tag{53}$$

A common application of this exponential distribution is the description of phenomena occuring independently at a constant rate, such as decay lengths and lifetimes. In view of the self-similar feature of the exponential function:

$$P(t - t_0 | t > t_0) \;=\; P(t)\,, \tag{54}$$

the exponential distribution is sometimes said to be memoryless.

4.2.3 The Gaussian distribution

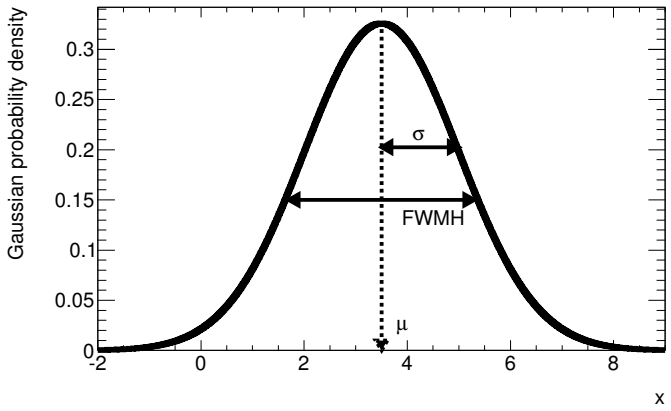


Fig. 7: A Gaussian PDF, with parameter values $\mu = 3.5$ and $\sigma = 1.5$.

Now turn to the Normal (or Gaussian) distribution. Consider a random variable x, with probability density

$$P_{\text{Gauss}}\,(x;\mu,\sigma) \;=\; \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,, \tag{55}$$

and with mean value and variance given by

$$E\,[x] \;=\; \mu\,, \tag{56}$$
$$V\,[x] \;=\; \sigma\,. \tag{57}$$

The PDF corresponding to the special $\mu = 0$, $\sigma = 1$ case is usually called "reduced normal".

On purpose, the same symbols μ and σ have been used both for the parameters of the Gaussian PDF and for the mean value and variance. This is an important feature: the Gaussian distribution is uniquely characterized by its first and second moments. For all Gaussians, the $[\mu - \sigma; \mu + \sigma]$ covers

68.3% of PDF integral, and is customary called a "one-sigma interval"; similarly for the two-sigma interval and its 95.4%.

The dispersion of a peaked distribution is sometimes characterised in terms of its FWHM (full width at half-maximum); for Gaussian distributions, this quantity is uniquely related to its variance, as $\mathrm{FWHM} = 2\sqrt{2\ln 2} \simeq 2.35\sigma$; Figure 7 provides a graphical illustration of the Gaussian PDF and its parameters.

In terms of conceptual relevance and practical applications, the Gaussian certainly outnumbers all other common distributions; this feature is largely due to the central limit theorem, which asserts that Gaussian distributions are the limit of processes arising from multiple random fluctuations. Consider n independent random variables $\vec{x} = \{x_1, x_2, \ldots, x_n\}$, each with mean and variances μ_i and σ_i^2; the variable $S(\vec{x})$, built as the sum of reduced variables

$$S \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu_i}{\sigma_i} \;, \tag{58}$$

can be shown to have a distribution that, in the large-n limit, converges to a reduced normal distribution, as illustrated in Figure 8 for the sum of up to five uniform distributions. Not surprisingly, any measurement subject to multiple sources of fluctuations is likely to follow a distribution that can be approximated with a Gaussian distribution to a good approximation, regardless of the specific details of the processes at play.

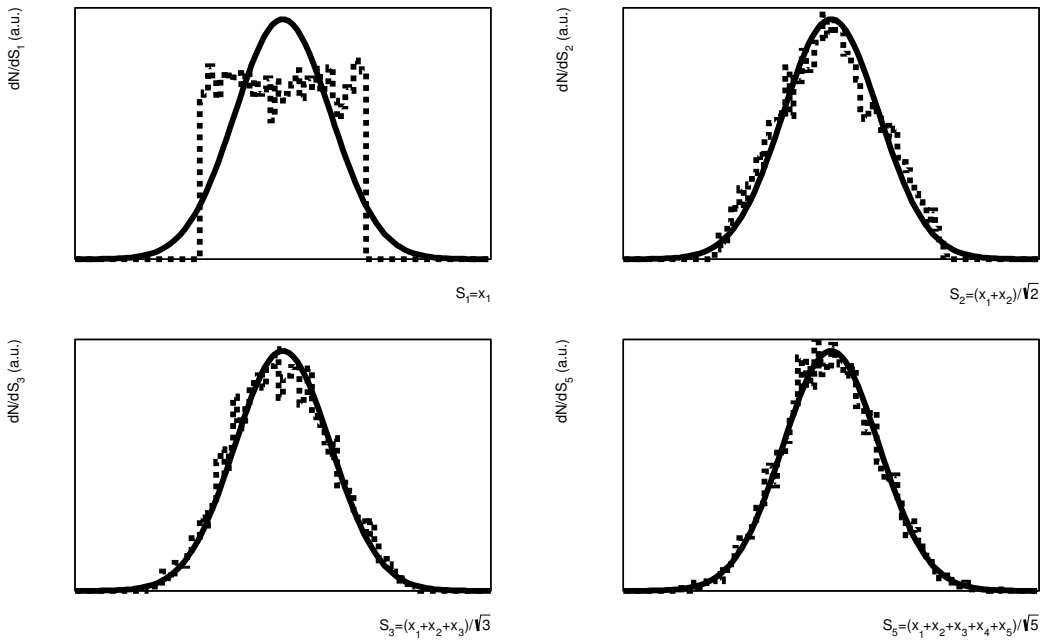


Fig. 8: A graphical illustration of the central limit theorem. The top left plot compares the histogram (dashed curve) of a sample realization from a uniform variable x_1, with a normal PDF (solid line) of same mean and variance; similarly, the plots on the top right, bottom left and bottom right plot the corresponding histograms for the sums S_2, S_3 and S_5 of two, three and five reduced uniform variables $x_1, \ldots, x_5$, respectively. The sequence of variables follow distributions that quickly converge to a reduced Gaussian.

The Gaussian is also encountered as the limiting distribution for the Binomial and and Poisson ones, in the large n and large λ limits, respectively:

$$P_{\mathrm{binomial}}\left(k; n \to \infty, p\right) \quad \to \quad P_{\mathrm{Gauss}}\left(k; np, np(1-p)\right) \;, \tag{59}$$

$$P_{\text{Poisson}}\left(k;\lambda\to\infty\right) \;\;\to\;\; P_{\text{Gauss}}\left(k;\lambda,\sqrt{\lambda}\right)\;. \tag{60}$$

Note that, when using a Gaussian as approximation, an appropriate continuity correction needs to be taken into account: the range of the Gaussian extends to negative values, while Binomial and Poisson are only defined in the positive range.

4.2.4 The χ^2 distributions

Now the χ^2 (or chi-squared) distributions are considered. The following PDF

$$P_{\chi^2}\left(x;n\right) \;=\; \begin{cases} \frac{x^{n/2-1}e^{-x/2}}{2^{n/2-1}\Gamma\left(\frac{n}{2}\right)} & ,\quad x\geq 0\,, \\ 0 & ,\quad \text{otherwise}\,, \end{cases} \tag{61}$$

with a single parameter n, and where $\Gamma\left(n/2\right)$ denotes the Gamma function, has mean value and variance given by

$$\begin{aligned} E\left[x\right] &\;=\; n\,, \\ V\left[x\right] &\;=\; 2n\,. \end{aligned} \tag{62}$$

The shape of the χ^2 distribution depends thus on n, as shown in Figure 9 where the χ^2 PDF, for the first six integer values of the n parameter, are shown. It can be shown that the χ^2 distribution can be written

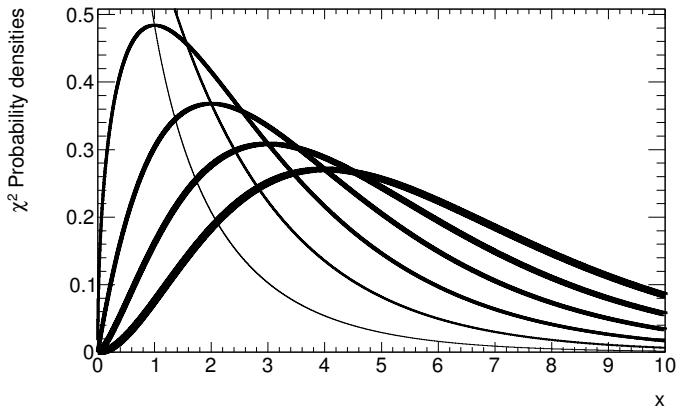


Fig. 9: The first six χ^2 probability density functions, for integer numbers of degrees of freedom n. The width of solid lines increases monotonically with n in the $n=1,\ldots,6$ range.

as the sum of squares of n normal-reduced variables x_i, each with mean μ_i and variance σ_i^2:

$$P_{\chi^2}\left(x;n\right) \;=\; \sum_{i=1}^{n}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\;. \tag{63}$$

In view of this important feature of the χ^2 distribution, the quantity n is called "number of degrees of freedom"; this name refers to the expected behaviour of a least-square fit, where n_d data points are used to estimate n_p parameters; the corresponding number of degrees of freedom is n_d-n_p. For a well-behaved fit, the χ^2 value should follow a χ^2 distribution. As discussed in 7.1, the comparison of an observed χ^2 value with its expectation, is an example of goodness-of-fit test.

4.2.5 The Breit-Wigner and Voigtian distributions

The survey closes with the discussion of two physics-motivated PDFs. The first is the Breit-Wigner function, which is defined as

$$P_{\mathrm{BW}}\left(x;\Gamma,x_0\right) \;=\; \frac{1}{\pi}\frac{\Gamma/2}{\left(x-x_0\right)^2+\Gamma^2/4}\;, \tag{64}$$

whose parameters are the most probable value x_0 (which specifies the peak of the distribution), and the FWHM Γ. The Breit-Wigner is also called Lorentzian by physicists, and in matematics it is often referred to as the Cauchy distribution. It has a peculiar feature, as a consequence of its long-range tails: the empirical average and empirical RMS are ill-defined (their variance increase with the size of the samples), and cannot be used as estimators of the Breit-Wigner parameters. The truncated mean and interquartile range, which are obtained by removing events in the low and high ends of the sample, are safer estimators of the Breit-Wigner parameters.

In the context of relativistic kinematics, the Breit-Wigner function provides a good description of a resonant process (for example the invariant mass of decay products from a resonant intermediate state); for a resonance, the parameters x_0 and Γ are referred to as its mass and its natural width, respectively.

Finally, the Voigtian function is the convolution of a Breit-Wigner with a Gaussian,

$$P_{\mathrm{Voigt}}\left(x;x_0,\Gamma,\sigma\right) \;=\; \int_{-\infty}^{+\infty} dx' P_{\mathrm{Gauss}}\left(x';0,\sigma\right) P_{\mathrm{BW}}\left(x-x';x_0,\Gamma\right)\;, \tag{65}$$

and is thus a three-parameter distribution: mass x_0, natural width Γ and resolution σ. While there is no straightforward analytical form for the Voigtian, efficient numerical implementations are available, i.e. the `TMath::Voigt` member function in the ROOT [7] data analysis framework, and the `RooVoigtian` class in the `RooFit` [8] toolkit for data modeling. For values of Γ and σ sufficiently similar, the FWHM of a Voigtian can be approximated as a combination of direct sums and sums in quadrature of the Γ and σ parameters. A simple, crude approximation yields :

$$\mathrm{FWHM}_{\mathrm{Voigt}} \;\simeq [(\Gamma/2)\oplus\sigma]+\Gamma/2\;. \tag{66}$$

When the instrumental resolutions are sufficiently well described by a Gaussian, a Voigtian distribution is a good model for observed experimental distributions of a resonant process. Figure 10 represents an analytical Breit-Wigner PDF (evaluated using the Z boson mass and width as parameters), and a dimuon invariant mass spectrum around the Z boson mass peak, as measured by the ATLAS experiment using 8 TeV data [9]. The width of the observed peak is interpreted in terms of experimental resolution effects, as indicated by the good data/simulation agreement. Of course, the ATLAS dimuon mass resolution is more complicated than a simple Gaussian (as hinted by the slightly asymmetric shape of the distribution), therefore a simple Voigtian function would not have reached the level of accuracy provided by the complete simulation of the detector resolution.

5 The maximum likelihood theorem and its applications

The discussion in Section 3.3 was aimed at describing a few intuitive examples of parameter estimation and their properties. Obviously, such case-by-case approaches are not general enough. The maximum likelihood estimation (MLE) is an important general method for parameter estimation, and is based on properties following the maximum likelihood (ML) theorem.

Consider a sample made of N independent outcomes of random variables $\vec{x}$, arising from a n-parametric PDF $P\left(\vec{x};\theta_1,\ldots,\theta_n\right)$, and whose analytical dependence on variables and parameters is known, but for which the value(s) of at least one of its parameters θ is unknown. With the MLE, these values are extracted from an analytical expression, called the likelihood function, that has a functional

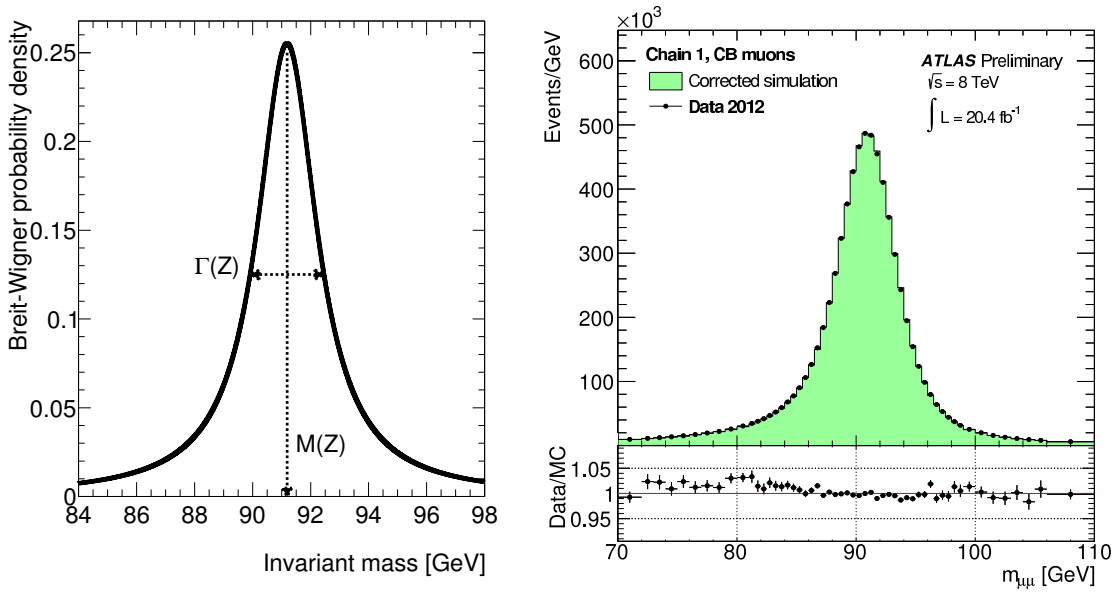


Fig. 10: Left: the probability density function for a Breit-Wigner distribution, using the PDG [6] values for the Z boson mass and width as parameters. Right: the invariant dimuon mass spectrum around the Z boson mass peak; the figure is from an ATLAS measurement using 8 TeV data [9].

dependence derived from the PDF, and is designed to maximise the probability of realizing the observed outcomes. The likelihood function $\mathcal{L}$ can be written as

$$\mathcal{L}\left(\theta_1, \ldots, \theta_n; \vec{x}\right) \; = \; \prod_{i=1}^{N} P\left(\vec{x}_i; \theta_1, \ldots, \theta_n\right) \; . \tag{67}$$

This notation shows implictly the functional dependence of the likelihood on the parameters θ, and on the N realizations $\vec{x}$ of the sample. The ML theorem states that the $\hat{\theta_1}, \ldots, \theta_n = \hat{\theta}_n$ values that maximize $\mathcal{L}$,

$$\mathcal{L}\left(\hat{\theta}_1, \ldots, \hat{\theta}_n \vec{x}\right) \; = \; \max|_\theta \left\{\mathcal{L}\left(\theta_1, \ldots, \theta_n\right)\right\} \; . \tag{68}$$

are estimators of the unknown parameters θ, with variances $\hat{\sigma}_\theta$ that are extracted from the covariance of $\mathcal{L}$ around its maximum.

In a few cases, the MLE can be solved analytically. A classical example is the estimation of the mean value and variance of an arbitrary sample, that can be analytically derived under the assumption that the underlying PDF is a Gaussian. Most often though, a numerical study of the likelihood around the θ parameter space is needed to localize the $\hat{\theta}$ point that minimizes $-\ln \mathcal{L}$ (the "negative log-likelihood", or NLL); this procedure is called a ML fit.

5.1 Likelihood contours

Formally speaking, several conditions are required for the ML theorem to hold. For instance, $\mathcal{L}$ has to be at least twice derivable with respect to all its θ parameters; constraints on (asymptotical) unbiased behaviour and efficiency must be satisfied; and the shape of $\mathcal{L}$ around its maximum must be normally distributed. This last condition is particularly relevant, as it ensures the accurate extraction of errors. When it holds, the likelihood is said (in a slightly abusive manner) to have a "parabolic" shape (more in

reference to $-\ln L$ than to the likelihood itself), and its expansion around $\hat{\theta}$ can be written as

$$f\left(\vec{\hat{\theta}}, \vec{\theta}, \mathbf{\Sigma}\right) \;=\; \frac{1}{\sqrt{2\pi}\,|\mathbf{\Sigma}|} \exp\left\{-\frac{1}{2}\left(\vec{\hat{\theta}}_i - \vec{\theta}_i\right) \Sigma_{ij}^{-1} \left(\vec{\hat{\theta}}_j - \vec{\theta}_j\right)\right\} . \tag{69}$$

In Eq. (69) the covariance matrix $\mathbf{\Sigma}$ has been introduced, and its elements are given by :

$$\mathbf{\Sigma}_{ij}^{-1} \;=\; -E\left[\frac{\partial \ln \mathcal{L}}{\partial \theta_i} \frac{\partial \ln \mathcal{L}}{\partial \theta_j}\right] . \tag{70}$$

When moving away from its maximum, the likelihood value decreases by amounts that depend on the covariance matrix elements:

$$-2\Delta \ln \mathcal{L} \;=\; -2\left[\ln \mathcal{L}(\vec{\theta}) - \ln \mathcal{L}(\vec{\hat{\theta}})\right] \;=\; \sum_{i,j} \left(\theta_i - \hat{\theta}_i\right) \mathbf{\Sigma}_{ij}^{-1} \left(\theta_i - \hat{\theta}_j\right) . \tag{71}$$

This result, together with the result on error propagation in Eq. (37), indicates that the covariance matrix defines contour maps around the ML point, corresponding to confidence intervals. In the case of a single-parameter likelihood $\mathcal{L}(\theta)$, the interval contained in a $-2\Delta \ln \mathcal{L} < 1$ contour around $\hat{\theta}$ defines a 68% confidence interval, corresponding to a $-\sigma_\theta \leq \theta - \hat{\theta} \leq \sigma_\theta$ range around the ML point; in consequence the result of this MLE can be quoted as $\left(\hat{\theta} \pm \hat{\sigma}_\theta\right)$.

5.2 Selected topics on maximum likelihood

5.2.1 Samples composed of multiple species

In a typical scenario, the outcome of a random process may arise from multiple sources. To be specific, consider that events in the data sample are a composition of two "species", called generically "signal" and "background" (the generalization to scenarios with more than two species is straighforward). Each species is supposed to be realized from its own probability densities, yielding similar (but not identical) signatures in terms of the random variables in the data sample; it is precisely these residual differences in PDF shapes that are used by the ML fit for a statistical separation of the sample into species. In the two-species example, the underlying total PDF is a combination of both signal and background PDFs, and the corresponding likelihood function is given by

$$\mathcal{L}\left(\theta; \vec{x}\right) \;=\; \prod_{i=1}^{N} \left[f_{\rm sig} P_{\rm sig}(\vec{x}; \theta) + (1 - f_{\rm sig}) P_{\rm bkg}(\vec{x}; \theta)\right] , \tag{72}$$

where $P_{\rm sig}$ and $P_{\rm bkg}$ are the PDFs for the signal and background species, respectively, and the signal fraction $f_{\rm sig}$ is the parameter quantifying the signal purity in the sample: $0 \leq f_{\rm sig} \leq 1$. Note that, since both $P_{\rm sig}$ and $P_{\rm bkg}$ satisfy the PDF normalization condition from Eq. (9), the total PDF used in Eq. (72) is also normalized. It is worth mentioning that some of the parameters θ can be common to both signal and background PDFs, and others may be specific to a single species. Then, depending on the process and the study under consideration, the signal fraction can either have a known value, or belong to the set of unknown parameters θ to be estimated in a ML fit.

5.2.2 Extended ML fits

In event-counting experiments, the actual number of observed events of a given species is a quantity of interest; it is then convenient to treat the number of events as an additional parameter λ of the likelihood function. In the case of a single species, this amounts to "extending" the likelihod,

$$\mathcal{L}\left(\lambda, \theta; \vec{x}\right) \;=\; \frac{\lambda^N e^{-\lambda}}{N!} \prod_{i=1}^{N} P\left(\vec{x_i}; \theta\right) . \tag{73}$$

where an additional multiplicative term, corresponding to the Poisson distribution (c.f. Section 4.1), has been introduced. (the $N!$ term in the denominator can be safely dropped; a global factor has no impact on the shape of the likelihood function nor on the ML fit results). It is straightforward to verify that the Poisson likelihod in Eq. (73) is maximal when $\hat{\lambda} = N$, as intended; now, if some of the PDFs also depend on λ, the value of $\hat{\lambda}$ that maximises $\mathcal{L}$ may differ from N. The generalization to more than one species is straightforward as well; for each species, a multiplicative Poisson term is included in the extended likelihood, and the PDFs of each species are weighted by their corresponding relative event fractions; in presence of two species, the extended version of the likelihood in Eq. 72 becomes:

$$\mathcal{L}\left(N_{\rm sig}, N_{\rm bkg}\theta; \vec{x}\right) \;=\; (N_{\rm sig} + N_{\rm bkg})^N e^{-(N_{\rm sig}+N_{\rm bkg})} \prod_{i=1}^{N} \left[N_{\rm sig} P_{\rm sig}(\vec{x};\theta) + N_{\rm bkg} P_{\rm bkg}(\vec{x};\theta)\right] \;. \qquad (74)$$

5.2.3 "Non-parabolic" likelihoods, likelihoods with multiple maxima

As discussed in Section 5.1, the condition in Eq. (69) about the asymptotically normal distribution of the likelihood around its maximum is crucial to ensure a proper interpretation of $-2\Delta \ln \mathcal{L}$ contours in terms of confidence intervals. In this paragraph, two scenarios in which this condition can break down are discussed.

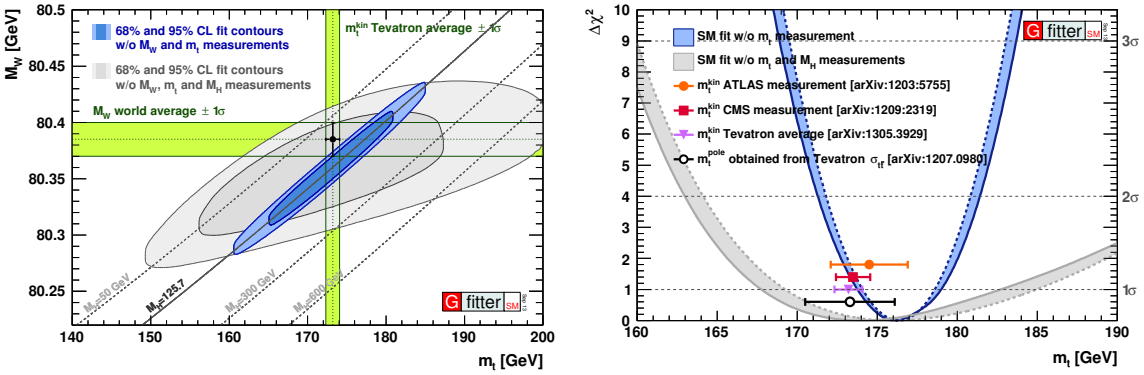


Fig. 11: Left: two-dimensional likelihood contours in the $m_w - m_t$ plane, obtained from a global electroweak fit (EW) performed by the gFitter collaboration [10]. Center: the change in the EW likelihood as a function of m_t, expressed in terms of $\delta\chi^2 = -2\Delta \ln L$. Both plots illustrate the non-parabolic shape of the EW likelihood.

A first scenario concerns a likelihood that is not totally symmetric around its maximum. Such a feature may occur, when studying low-statistics data samples, in view of the Binomial or Poissonian behaviour of event-counting related quantities (c.f. Figure 5). But it can also occur on larger-sized data samples, indicating that the model has a limited sensitivity to the parameter θ being estimated, or as a consequence of strong non-linear relations in the likelihood function. As illustration, examples of two- and one-dimensional likelihood contours, with clear non-parabolic shapes, are shown in Figure 11.

Also, the likelihood function may possess one or more local maxima, or even completely degenerate maxima. There are various possible sources for such degeneracies. For example in models with various species, if some PDF shapes are not different enough to provide a robust statistical discrimination among their corresponding species. For example, if swapping the PDFs for a pair of species yields a sufficiently good ML fit, a local maximum may emerge; on different sample realizations, the roles of local and global maxima may alternate among the correct and swapped combinations. The degeneracies could also arise as a reflexion of explicit, physical symmetries in the model: for example, time-dependent asymmetries in $B^0 \to \pi^+\pi^-$ decays are sensitive to the CKM angle α, but the physical observable is a function of $\cos 2(\alpha + \delta)$, with an addtional phase δ at play; in consequence, the model brings up to eight

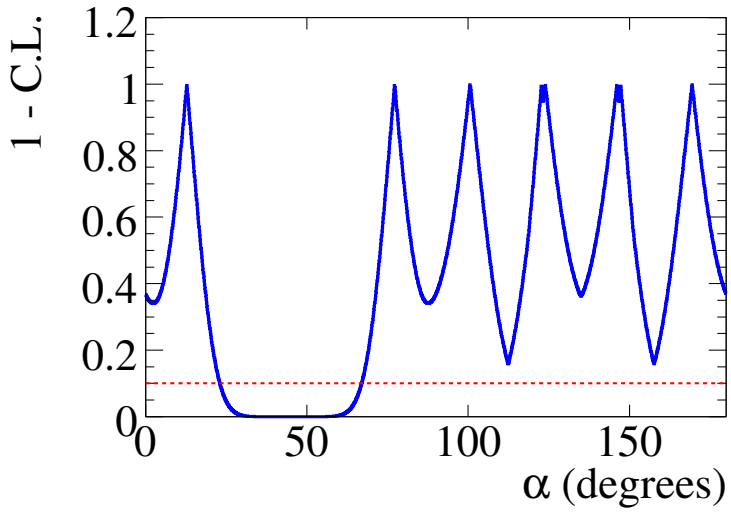


Fig. 12: The change in likelihood, expressed in terms of a variant called "confidence level" CL, shown as a function of the CKM angle α; the definition of CL is such that $\mathrm{CL} = 0$ at the solutions of the ML fit. An eight-fold constraint on α is extracted from measurements in $B \to \pi\pi$ decays by the *BABAR* experiment [11]. The two pairs of solutions around $\sim 130^\circ$ are very close in values, and barely distinguishable in the figure.

indistinguishible solutions for the CKM angle α, as illustrated in Figure 12.

In all such cases, the (possibly disjoint) $-2\Delta \ln \mathcal{L} < 1$ interval(s) around the $\hat{\theta}$ central value(s) cannot be simply reported with a symmetric uncertainty $\hat{\sigma}_\theta$ only. In presence of a single, asymmetric solution, the measurement can be quoted with asymmetric errors, i.e. $\hat{\theta}^{+\sigma_+}_{-\sigma_-}$, or better yet, by providing the detailed shape of $-2\Delta \ln \mathcal{L}$ as a function of the estimated parameter θ. For multiple solutions, more information needs to be provided: for example, amplitude analyses ("Dalitz-plot" analyses) produced by the B-factories *BABAR* and Belle, often reported the complete covariance matrices around each local solution (see e.g. [12, 13] as examples).

5.2.4 Estimating efficiencies with ML fits

Consider a process with two possible outcomes: "yes" and "no". The intuitive estimator of the efficiency ε is a simple ratio, expressed in terms of the number of outcomes n_{yes} and n_{no} of each kind:

$$\hat{\varepsilon} \;=\; \frac{n_{\mathrm{yes}}}{n_{\mathrm{yes}} + n_{\mathrm{no}}} \,, \tag{75}$$

for which the variance is given by

$$V\left[\hat{\varepsilon}\right] \;=\; \frac{\hat{\varepsilon}(1-\hat{\varepsilon})}{n} \,, \tag{76}$$

where $n = n_{\mathrm{yes}} + n_{\mathrm{no}}$ is the total number of outcomes realized. This estimator $\hat{\varepsilon}$ clearly breaks down for low n, and in the very low or very high efficiency regimes. The MLE technique offers a robust approach to estimate efficiencies: consider a PDF $P(x;\theta)$ to model the sample, and include in it an additional discrete, bivariate random variable $c = \{\mathrm{yes}, \mathrm{no}\}$, so that the PDF becomes

$$P\left(x, c; \theta\right) \;=\; \delta(c - \mathrm{yes})\varepsilon(x,\theta) + \delta(c - \mathrm{no})\left[1 - \varepsilon(x,\theta)\right] \,. \tag{77}$$

In this way, the efficiency is no longer a single number, but a function of x (plus some parameters θ that may be needed to characterize its shape). With this function, the efficiency can be extracted in different x domains, as illustated in Figure 13, or can be used to produce a multidimensional efficiency map.

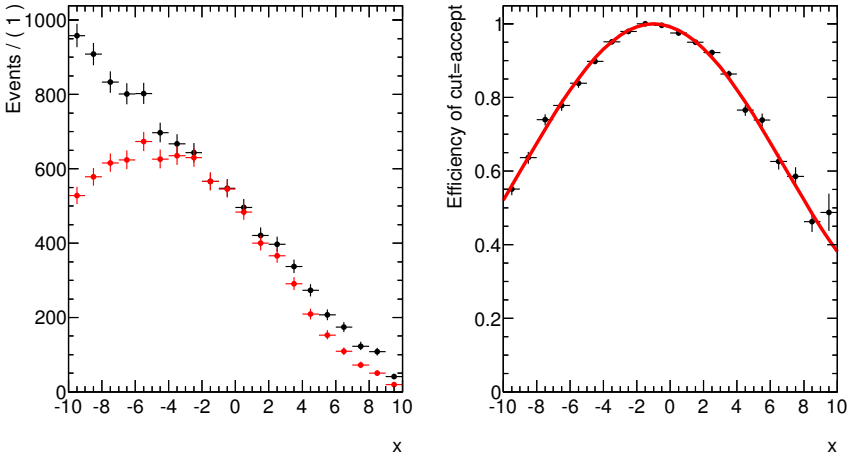


Fig. 13: Left: the frequency of events in a sample, as a function of a variable x. The sample contains two categories of events: "accepted" and "rejected". The red dots indicate the bins of the histogram for those "accepted" only, and the black dots for the two categories cumulated. Right: the efficiency as a parametric function of x, with parameter values extracted from a ML fit to the data sample. Figure extracted from the `RooFit` user's guide [8].

5.3 Systematic uncertainties

As discussed in Section 5.1, in MLE the covariance of $\mathcal{L}$ is the estimator of statistical uncertainties. Other potential sources of (systematical) uncertainties are usually at play, and need to be quantified. For this discussion, a slight change in the notation with respect to Eq.(67) is useful; in this notation, the likelihood function is now written as

$$\mathcal{L}\left(\mu_1, \ldots, \mu_p \ , \ \theta_1, \ldots, \theta_k; \vec{x}\right) \ , \tag{78}$$

where the parameters are explicitly partitioned into a set $\mu_1, \ldots, \mu_p$, called parameters of interest (POI), that correspond to the actual quantities that are to be estimated; and a set $\theta_1, \ldots, \theta_k$, called nuisance parameters (NP), that represent potential sources of systematic biases: if inaccurate or wrong values are assigned to some NPs, the shapes of the PDFs can be distorted, the estimators of POIs can become biased. The systematic uncertainties due to NPs are usually classified in two categories.

The first category refers to "Type-I" errors, for which the sample (or other control data samples) can (in principle) provide information on the NPs under consideration, and are (in principle) supposed to decrease with sample size. The second category refers to "Type-II" errors, which arise from incorrect assumptions in the model (i.e. a choice of inadequate functional dependences in the PDFs), or uncontrolled features in data that can not be described by the model, like the presence of unaccounted species. Clearly, for Type-II errors the task of assigning systematic uncertainties to them may not be well-defined, and may spoil the statistical interpretation of errors in termos of confidence levels.

5.3.1 The profile-likelihod method

To deal with Type-I nuisance parameters, a popular approach is to use the so-called profile-likelihood method. This approach consists of assigning a specific likelihood to the nuisance parameters, so that the original likelihood is modified to have two different components:

$$\mathcal{L}\left(\mu, \theta\right) \ = \ \mathcal{L}_\mu\left(\mu, \theta\right) \mathcal{L}_\theta\left(\theta\right) \ . \tag{79}$$

Then, for a fixed value of μ, the likelihood is maximized with respect to the nuisance θ; the sequencial outcome of this procedure, called profile likelihood, is a function that depends only on μ: it is then said

that the nuisance parameter has been profiled out of the likelihood.

As an example, consider the measurement of the cross-section of a generic process, $\sigma\,(\mathrm{initial} \to \mathrm{final})$ If only a fraction of the processes is actually detected, the efficiency ε of reconstructing the final state is needed to convert the observed event rate $\hat{N}_{\mathrm{event}}$ into a measurement $\hat{\sigma}$. This efficiency is clearly a nuisance: a wrong value of ε directly affects the value of $\hat{\sigma}$, regardless of how accurately $\hat{N}_{\mathrm{event}}$ may have been measured. By estimating $\hat{\varepsilon}$ on a quality control sample (for example, high-statistics simulation, or a high-purity control data sample), the impact of this nuisance can be attenuated. For example, an elegant analysis would produce a simultaneous fit to the data and control samples, so that the values and uncertainties of NPs are estimated in the ML fit, and are correctly propagated to the values and variances of the POIs.

As another example, consider the search for a resonance ("a bump") over a uniform background. If the signal fraction is very small, the width Γ of the bump cannot be directly estimated on data, and the value used in the signal PDF has to be inferred from external sources. This width is clearly a nuisance: using an overestimated value would translate into an underestimation of the signal-to-background ratio, and thus an increase in the variance of the signal POIs, and possibly biases in their central values as well, i.e. the signal rate would tend to be overestimated. Similar considerations can be applied in case of underestimation of the width. If an estimation $\hat{\Gamma} \pm \hat{\sigma}_{\Gamma}$ of the width is available, this information can be implemented as in Eq. (79), by using a Gaussian PDF, with mean value $\hat{\Gamma}$ and width $\hat{\Gamma}$, in the $\mathcal{L}_{\Gamma}$ component of the likelihood. This term acts as a penalty in the ML fit, and thus constraints the impact of the nuisance Γ on the POIs.

6 Multivariate techniques

Often, there are large regions in sample space where backgrounds are overwhelming, and/or signals are absent. By restricting the data sample to "signal-enriched" subsets of the complete space, the loss of information may be minimal, and other advantages may compensate the potential losses: in particular for multi-dimensional samples, it can be difficult to characterize the shapes in regions away from the core, where the event densities are low; also reducing the sample size can relieve speed and memory consumption in numerical computations.

The simplest method of sample reduction is by requiring a set of variables to be restricted into finite intervals. In practice, such "cut-based" selections appear at many levels in the definition of sample space: thresholds on online trigger decisions, filters at various levels of data acquisition, removal of data failing quality criteria... But at more advanced stages of a data analysis, such "accept-reject" sharp selections may have to be replaced by more sophisticated procedures, generically called multivariate techniques.

A multi-dimensional ML fit is an example of a multivariate technique. For a MLE to be considered, a key requirement is to ensure a robust knowledge of all PDFs over the space of random variables.

Consider a set of n random variables $\vec{x} = \{x_1, x_2 \ldots, x_n\}$. If all variables are shown to be uncorrelated, their n-dimensional PDF is completely determined by the product of their n one-dimensional PDFs; now, if variables are correlated, but their correlation patterns are completely linear, one can instead use variables $\vec{y}$, linear combinations of $\vec{x}$ obtained by diagonalizing the inverse covariance. For some non-linear correlation patterns, it may be possible to find analytical descriptions; for instance, the (mildly) non-linear correlation pattern represented in Figure 2, was produced with the `RooFit` package, by applying the `Conditional` option in `RooProdPdf` to build a product of PDFs. In practice, this elegant approach cannot be easily extended to more than two dimensions, and is not guaranteed to reproduce complex, non-linear patterns. In such scenarios, the approach of dimensional reduction can potentially bring more effective results.

A typical scenario for dimensional reduction is when several variables carry common information (and thus exhibit strong correlations), together with some diluted (but relevant) pieces of independent in-

formation. An example is the characterization of showers in calorimeters; for detectors with good transverse and/or longitudinal segmentation, the signals deposited in nearby calorimetric channels can be used to reconstruct details from the shower development; for example, a function that combines informations from longitudinal and transverse shower shapes, can be used to discriminate among electromagnetic and hadronic showers.

The simplest algorithm for dimensional reduction is the Fisher discriminant: it is a linear function of variables, with coefficients adjusted to match an optimal criterion, called separation among two species, which is is the ratio of the variance between the species to the variance within the species, and can be expressed in a close, simple analytical form.

In presence of more complex, non-linear correlation patterns, a large variety of techniques and tools are available. The TMVA [14] package is a popular implementation of dimensional-reduction algorithms; other than linear and likelihood-based discriminants, it provides easy training and testing methods for artificial neural networks and (boosted) decision trees, which are among those most often encountered in HEP analyses. As a general rule, a multivariate analyzer uses a collection of variables, realized on two different samples (corresponding to "signal" and "background" species), to perform a method-dependent training, guided by some optimization criteria; then performances of the trained analyzer are evaluated on independent realizations of the species (this distinction between the training and testing stages is crucial to avoid "over-training" effects). Figure 14 shows a graphical representation of a figure-of-merit comparison of various analyzers implemented in TMVA.

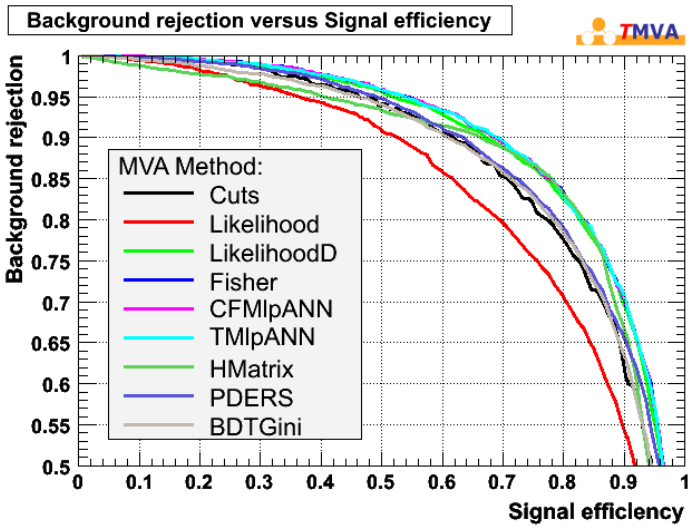


Fig. 14: A figure-of-merit comparison between various different multivariate analyzer methods using the TMVA package. The training was performed on two simulated samples ("signal" and "background"), each consisting of four linearly correlated Gaussian-distributed variables. The various lines indicate the trade-off between signal efficiency and background rejection. The figure is taken from [14].

7 Statistical hypothesis testing

Sections 3 and 5 mainly discussed procedures for extracting numerical information from data samples; that is, to perform measurements and report them in terms of values and uncertanties. The next step in data analysis is to extract qualitative information from data: this is the domain of statistical hypothesis testing. The analytical tool to assess the agreement of an hypothesis with observations is called a test statistic, (or a statistic in short); the outcome of a test is given in terms of a p-value, or probability of a "worse" agreement than the one actually observed.

7.1 The χ^2 test

For a set of n independent measurements x_i, their deviation with respect to predicted values μ_i, expressed in units of their variances σ_i is called the χ^2 test, and is defined as

$$\chi^2 \;=\; \sum_{\mathrm{i}=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \;. \tag{80}$$

An ensemble of χ^2 tests is a random variable that, as mentioned in Section 4.2, follows the P_{χ^2} distribution (c.f. Eq. 61) for n degrees of freedom. Its expectation value is n and its variance $2n$; therefore one does expect the χ^2 value observed in a given test not to deviate much from the number of degrees of freedom, and so this value can be used to probe the agreement of prediction and observation. More precisely, one expects 68% of tests to be contained within a $n \pm \sqrt{2n}$ interval, and the p-value, or probability of having a test with values larger than a given χ^2 value is

$$p \;=\; \int_{\chi^2}^{+\infty} dq \; P_{\chi^2}\left(q; n\right) \;. \tag{81}$$

Roughly speaking, one would tend to be suspicious of small observed p-values, as they may indicate a trouble, either with the prediction or the data quality. The interpretation of the observed p-value (i.e. to decide whether it is too small or large enough) is an important topic, and is discussed below in a more general approach.

7.2 General properties of hypothesis testing

Consider two mutually excluding hypotheses H_0 and H_1, that may describe some data sample; the hypothesis testing procedure states how robust is H_0 to describe the observed data, and how incompatible is H_1 with the observed data. The hypothesis H_0 being tested is called the "null" hypothesis, while H_1 is the "alternative" hypothesis,

Note that in the context of the search for a (yet unknown) signal, the null H_0 corresponds to a "background-only" scenario, and the alternative H_1 to "signal-plus-background"; while in the context of excluding a (supposedly inexistent) signal, the roles of the null and alternative hypotheses are reversed: the null H_0 is "signal-plus-background" and the alternative H_1 is "background-only".

The sequence of a generic test, aiming at accepting (or rejecting) the null hypothesis H_0 by confronting it to a data sample, can be sketched as follows:

- build a test statistic q, that is, a function that reduces a data sample to a single numerical value;
- define a confidence interval $W \;\rightarrow\; [q_{\mathrm{lo}} : q_{\mathrm{hi}}]$;
- measure $\hat{q}$;
- if $\hat{q}$ is contained in W, declare the null hypothesis accepted; otherwise, declare it rejected.

To characterize the outcome of this sequence, two criteria are defined: a "Type-I error" is incurred in, if H_0 is rejected despite being true; while a "Type-II error" is incurred in, if H_0 is accepted despite being false (and thus H_1 being true). The rates of Type-I and Type-II errors are called α and β respectively, and are determined by integrating the H_0 and H_1 probability densities over the confidence interal W,

$$\begin{aligned} 1 - \alpha \;&=\; \int_W dq \mathcal{P}\left(q|H_0\right) \;, \\ \beta \;&=\; \int_W dq \mathcal{P}\left(q|H_1\right) \;. \end{aligned} \tag{82}$$

The rate α is also called "size of the test" (or size, in short), as fixing α determines the size of the confidence interval W. Similarly, $1 - \beta$ is also called "power". Together, size and power characterize

the performance of a test statistic; the Neyman-Pearson lemma states that the optimal statistic is the likelihood ratio q_λ,

$$q_\lambda\left(\text{data}\right) \;=\; \frac{\mathcal{L}\left(\text{data}|H_0\right)}{\mathcal{L}\left(\text{data}|H_1\right)}\;. \tag{83}$$

The significance of the test is given by the p-value,

$$p \;=\; \int_{\hat{q}}^{+\infty} dq\mathcal{P}\left(q|H_0\right)\;. \tag{84}$$

which is often quoted in terms of "sigmas",

$$p \;=\; \int_{n\sigma}^{+\infty} dz\frac{1}{\sqrt{2\pi}}e^{-z^2/2} \;=\; 1-\frac{1}{2}\text{erf}\left(\frac{n}{\sqrt{2}}\right)\;, \tag{85}$$

so that for example a $p < 0.0228$ outcome can be reported as a "two-sigma" effect. Alternatively, it is common practice to quote the complement of the p-value as a confidence level (C.L.).

The definition of a p-value as in Eq. (84) (or similarly in Eq. (81) for the example for a χ^2 test) is clear and unambiguous. But interpretation of p-values is partly subjective: the convenience of a numerical "threshold of tolerance" may depend on the kind of hypothesis being tested, or on common practice. In HEP usage, three different traditional benchmarks are conventionally employed:

- in exclusion logic, a 95% C.L. threshold on a signal-plus-background test to claim exclusion;
- in discovery logic, a three-sigma threshold ($p < 1.35\times 10^{-3}$) on a background-only test to claim "evidence";
- and a five-sigma threshold ($p < 2.87\times 10^{-7}$) on the background-only test is required to reach the "observation" benchmark.

7.3 From LEP to LHC: statistics in particle physics

In experimental HEP, there is a tradition of reaching consensual agreement on the choices of test statistics. The goal is to ensure that, in the combination of results from different samples and instruments, the detector-related components (specific to each experiment) factor out from the physics-related observables (which are supposed to be universal). For example in the context of searches for the Standard Model (SM) Higgs boson, the four LEP experiments agreed on analyzing their data using the following template for their likelihoods:

$$\begin{aligned}
\mathcal{L}\left(H_1\right) &= \prod_{a=1}^{N_{\text{ch}}}\mathcal{P}_{\text{Poisson}}\left(n_a, s_a+b_a\right)\prod_{j=1}^{n_a}\frac{s_a\mathcal{S}_a(\vec{x_j})+b_a\mathcal{B}_a(\vec{x_j})}{s_a+b_a}\;,\\
\mathcal{L}\left(H_0\right) &= \prod_{a=1}^{N_{\text{ch}}}\mathcal{P}_{\text{Poisson}}\left(n_a, b_a\right)\prod_{j=1}^{n_a}\mathcal{B}_a(\vec{x_j})\;.
\end{aligned} \tag{86}$$

where N_{ch} is the number of Higgs decay channels studied, n_a is the observed number of event candidates in channel a, $\mathcal{S}_a$ and s_a ($\mathcal{B}_a$ and b_a) are the PDF and event yield for the signal (background) species in that channel. Also, the test statistic λ, derived from a likelihood ratio, is

$$\lambda \;=\; -2\ln Q\;,\;\text{with}\;Q \;=\; \frac{\mathcal{L}\left(H_1\right)}{\mathcal{L}\left(H_0\right)}\;; \tag{87}$$

so that roughly speaking, positive values of λ favour a "background-like" scenario, and negative ones are more in tune with a "signal-plus-background" scenario; values close to zero indicate poor sensitivity to distinguish among the two scenarios. The values use to test these two hypotheses are:

– under the background-only hypothesis, $\mathrm{CL}(b)$ is the probability of having a $-2\ln Q$ value larger than the observed one;
– under the signal+plus+background hypothesis, $\mathrm{CL}(s+b)$ is the probability of having a $-2\ln Q$ value larger than the observed one.

Figure 15 shows, for three different Higgs mass hypotheses, the $-2\ln Q$ values from the combination of the four LEP experiments in their searches for the SM Higgs boson, overlaid with the expected $\mathrm{CL}(s+b)$ and $1-\mathrm{CL}(b)$ distributions. Figure 16 shows the evolution of $-2\ln Q$ values as a function of the hypothetized Higgs boson mass; (as stated in the captions, the color conventions in the one- and two-dimensional plots are different) note that for masses below ~ 115 GeV, a positive value of the $-2\ln Q$ test statistic would have provided evidence for a signal; and sensitivity is quickly lost above that mass.

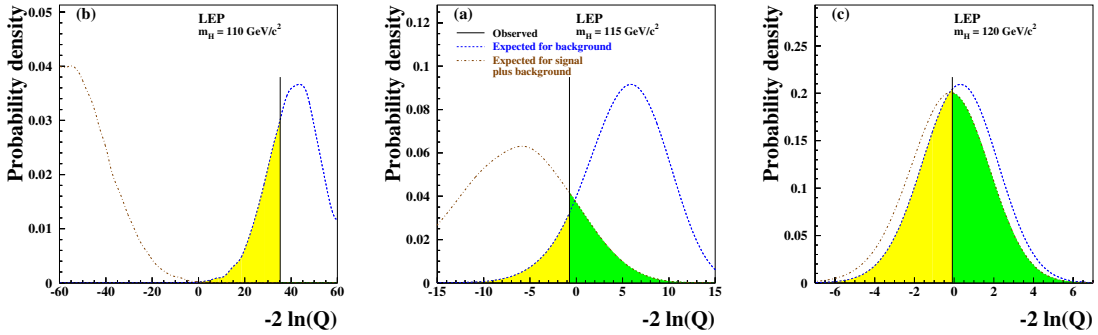


Fig. 15: Combined results from the search for a SM Higgs boson, performed by the four LEP experiments. From left to right, the $m_H = 110, 115, 120$ GeV hypotheses for the Higgs mass are used. The dashed blue and dashed red lines correspond to the PDFs for the background-only and signal-plus-background hypotheses, respectively. The observed values of the test statistic $-2\ln Q$ are marked by black vertical lines. The yellow areas indicate the $1-\mathrm{CL}(b)$ values for the background-only hypothesis, and the green areas the $\mathrm{CL}(s+b)$ value for signal-plus-background. The three figures are taken from [17].

7.3.1 The modified $\mathrm{CL}(s)$ *hypothesis testing*

The choice of $\mathrm{CL}(s+b)$ to test the signal-plus-background hypothesis, while suitably defined as a p-value. may drag some subjective concern: in case of a fortuitous simultaneous downward fluctuation in both signal and background, the standard 95% benchmark may lead to an exclusion of the signal, even if the sensitivity is poor.

A modification of the exclusion benchmark called $\mathrm{CL}(s)$, has been introduced in this spirit [18], and is defined as

$$\mathrm{CL}(s) \;=\; \frac{\mathrm{CL}(s+b)}{1-\mathrm{CL}(b)}\,. \tag{88}$$

This test, while not corresponding to a p-value (a ratio of probabilities is not a probability), has the desired property of protecting against downwards fluctuations of the background, and is commonly used in exclusion results, including the searches for the SM Higgs boson from the Tevatron and LHC experiments.

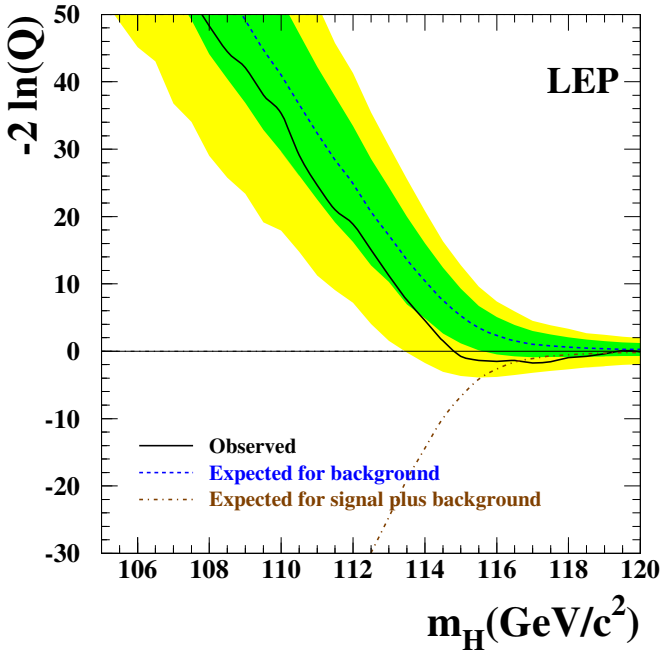


Fig. 16: The test statistic $-2\ln Q$ as a function of the Higgs boson mass m_H, obtained by combining the data of the four LEP experiments. The dashed line is the mean value of the background-only distribution at each mass m_H, and the green and yellow areas represent 68% and 95% contours for the background-only distribution at each mass m_H. The black line follows the $-2\ln Q$ value observed in data as a function of m_H. Figure taken from [17].

7.3.2 Profiled likelihood ratios

Following the recommendations from the LHC Higgs Combination Group [19], the ATLAS and CMS experiments have agreed on using a common test statistic, called profiled likelihood ratio and defined as

$$\tilde{q}_\mu\left(\mu\right) \;=\; -2\ln\frac{\mathcal{L}\left(\mu,\hat{\hat{\theta}}\right)}{\mathcal{L}\left(\hat{\mu},\hat{\theta}\right)}\;,\;\text{with}\;0\leq\hat{\mu}\leq\mu\;, \tag{89}$$

where the PIO $\mu = \sigma/\sigma_{\rm SM}$ is the "signal strength modifier", or Higgs signal rate expressed in units of the SM predicted rate, $\hat{\hat{\theta}}$ are the fitted values of the NPs at fixed values of the signal strength, and $\hat{\mu}$ and $\hat{\theta}$ are the fitted values when both μ and NPs are all free to vary in the ML fit [2]. The lower constraint on $0 \leq \hat{\mu} \leq \mu$ ensures that the signal rate is positive, and the upper constraint imposes that an upward fluctuation would not disfavor the SM signal hypothesis.

For an observed statistic value $\hat{\tilde{q}}_\mu$, the p-values for testing the signal-plus-background and background-only hypotheses, $p(s+b)$ and $p(b)$, are

$$p(s+b) \;=\; \int_{\hat{\tilde{q}}_\mu}^{\infty} dqP\left(q;\mu=\hat{\mu},\hat{\hat{\theta}}\right)\;, \tag{90}$$

$$1-p(b) \;=\; \int_{\hat{\tilde{q}}_\mu}^{\infty} dqP\left(q;\mu=0,\hat{\hat{\theta}}\right)\;. \tag{91}$$

[2] The test statistic actually reported in [19] is slightly different than the one described here, but in the context of these notes this subtlety can be disregarded.

and the results on searches are reported both in terms of the exclusion significance using the $\mathrm{CL}(s)$ observed and expected values, and the observation significance expressed in terms of the "local" [3] $p(b)$ expected and observed values.

To conclude the discussion on hypothesis testing, there is certainly no better illustration than Figure 17, taken from the results announced by the ATLAS and CMS experiments on July 4th, 2012: in the context of the search for the SM Higgs boson, both collaborations established the observation of a new particle with a mass around 125 GeV.

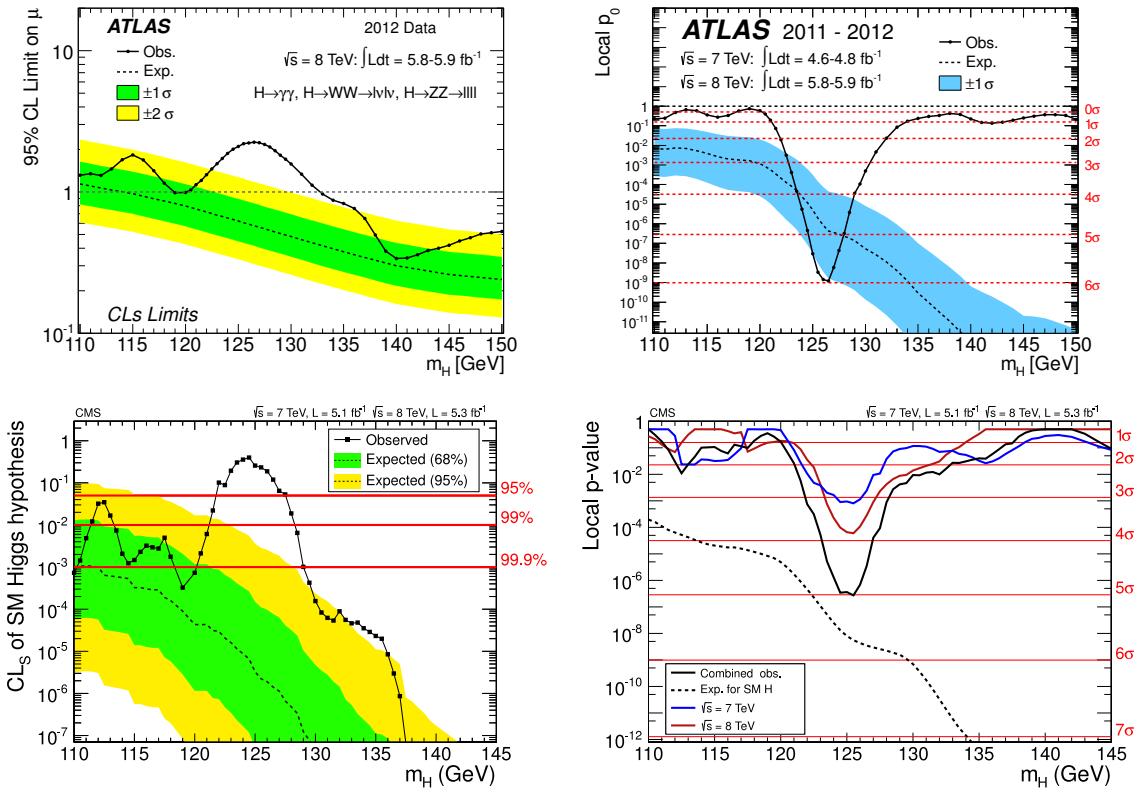


Fig. 17: The expected distributions and observed values for the exclusion (left) and the observation significances (right) in searches for the SM Higgs boson presented by the ATLAS [21] (top) and CMS [22] (bottom) collaborations on July 4th, 2012. Both experiments find a significant excess around 125 GeV.

8 Conclusions

These lectures aimed at providing a pedagogical overview of probability and statistics. The choice of topics was driven by practical considerations, based on the tools and concepts actually encountered in experimental high-energy physics. A bias in the choices, induced by the author's perspective and personal experience is not to be excluded. The discussion was completed with examples from recent results, mostly (although not exclusively) stemming from the B-factories and the LHC experiments.

[3] In short, the observation significance must be corrected for the trials factor, or "look-elsewhere effect"; in the case of the search for the SM Higgs boson in a wide Higgs mass interval, this effect translates into a decrease of the significance, as different Higgs masses are tested using independent data, and thus the probability of observing a signal-like fluctuation depends on the mass interval studied and the experimental resolution.

9 Acknowledgements

I would like to express my gratitude to all who helped make the AEPSHEP 2012 school a success: the organizers, the students, my fellow lecturers and discussion leaders. The highly stimulating atmosphere brought fruitful interactions during lectures, discussion sessions and conviviality activities.

References

[1] A. Stuart, J.K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2A: *Classical Inference and the Linear Model* 6th Ed., Oxford Univ. Press (1999), and earlier editions by Kendall and Stuart.

[2] R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989.

[3] G.D. Cowan, *Statistical Data Analysis*, Oxford University Press, 1998.

[4] F. James, *Statistical Methods in Experimental Physics*, 2nd ed., World Scientific, 2006.

[5] L. Lyons, *Statistics for Nuclear and Particle Physicists*, Cambridge University Press, 1986.

[6] J. Beringer *et al.* [Particle Data Group Collaboration], Phys. Rev. D **86**, 010001 (2012).

[7] R. Brun and F. Rademakers, Nucl. Instrum. Meth. A **389**, 81 (1997).

[8] W. Verkerke and D. P. Kirkby, eConf C **0303241**, MOLT007 (2003).

[9] The ATLAS collaboration, ATLAS-CONF-2013-088.

[10] M. Baak, M. Goebel, J. Haller, A. Hoecker, D. Kennedy, R. Kogler, K. Moenig and M. Schott *et al.*, Eur. Phys. J. C **72**, 2205 (2012).

[11] J. P. Lees [BaBar Collaboration], Phys. Rev. D **87**, no. 5, 052009 (2013).

[12] B. Aubert *et al.* [BaBar Collaboration], Phys. Rev. D **80** (2009) 112001.

[13] J. Dalseno *et al.* [Belle Collaboration], Phys. Rev. D **79**, 072004 (2009).

[14] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, "TMVA: Toolkit for Multivariate Data Analysis," PoS A CAT 040 (2007).

[15] K. Cranmer *et al.* [ROOT Collaboration], "HistFactory: A tool for creating statistical models for use with RooFit and RooStats," CERN-OPEN-2012-016.

[16] L. Moneta, K. Belasco, K. S. Cranmer, S. Kreiss, A. Lazzaro, D. Piparo, G. Schott and W. Verkerke *et al.*, The RooStats Project," PoS ACAT **2010**, 057 (2010).

[17] R. Barate *et al.* [LEP Working Group for Higgs boson searches and ALEPH and DELPHI and L3 and OPAL Collaborations], Phys. Lett. B **565**, 61 (2003).

[18] A. L. Read, J. Phys. G **28**, 2693 (2002).

[19] [ATLAS and CMS Collaborations], "Procedure for the LHC Higgs boson search combination in summer 2011," ATL-PHYS-PUB-2011-011, CMS-NOTE-2011-005.

[20] G. Cowan, K. Cranmer, E. Gross and O. Vitells, Eur. Phys. J. C **71**, 1554 (2011)

[21] G. Aad *et al.* [ATLAS Collaboration], Phys. Lett. B **716**, 1 (2012).

[22] S. Chatrchyan *et al.* [CMS Collaboration], Phys. Lett. B **716**, 30 (2012).

Local Organizing Committee

Takeo Higuchi (KEK)
Hiroyuki Iwasaki (KEK)
Kiyotomo Kawagoe (chair) (Kyushu University)
Takasumi Maruyama (KEK)
Mitsuaki Nozaki (KEK)
Ken-ichi Okumura (Kyushu University)
Susumu Oda (Kyushu University)
Akira Sugiyama (Saga University)
Motoi Tachibana (Saga University)
Junji Tojo (Kyushu University)
Katsuo Tokushuku (KEK)
Tamaki Yoshioka (Kyushu University)

International Organizing Committee

Mark Boland (Australian Synchrotron)
Subhasis Chattopadhyay (VECC)
Simon Eidelman (BINP)
Nick Ellis (chair) (CERN)
Kazunori Hanagaki (Osaka University)
Yee Bob Hsiung (National Taiwan University)
Pyungwon Ko (KIAS)
Sreerup Raychaudhuri (TIFR)
Lydia Roos (IN2P3/LPNHE Paris)
Didier Vilanova (Irfu/CEA)
Changzheng Yuan (IHEP)
Shi-Lin Zhu (Peking University)

International Advisory Committee

Etienne Augé (IN2P3/CNRS)
Mikhail Danilov (ITEP)
Rohini Godbole (Indian Institute of Science)
Shih-Chang Lee (Academia Sinica)
Martin Sevior (University of Melbourne)
Dongchul Son (Kyungpook National University)
Fumihiko Takasaki (chair) (KEK)
Rüdiger Voss (CERN)
Yifang Wang (IHEP)

Lecturers

Luis Alvarez-Gaume (CERN)
Shoji Asai (University of Tokyo)
Emi Kou (LAL Orsay)
Hai-Bo Li (IHEP)
Hsiang-Nan Li (Academia Sinica)
Hitoshi Murayama (IPMU, University of Tokyo and LBNL)
Mihoko Nojiri (KEK)
José Ocariz (IN2P3 and University Paris Diderot)
Werner Riegler (CERN)
Rajeev S. Bhalerao (TIFR)
Zhi-Zhong Xing (IHEP)

Discussion Leaders

Marc Besançon (Irfu/CEA)
John Ellis (King's College London and CERN)
Kin-ya Oda (Kyoto University)
Pavel Pakhlov (ITEP)
Myeonghun Park (CERN)
Francesco Riva (EPFL and IFAE)
Raymond Volkas (University of Melbourne)

Students

Curtis BLACK
Sabyasachi CHAKRABORTY
Emyr CLEMENT
Camille COUTURIER
Francesca DORDEI
Matteo FRANCHINI
Saranya Samik GHOSH
Philippe GROS
Mototsugu HACHIMINE
maryam HASHEMINIA
Simon HEISTERKAMP
Shigeki HIROSE
Bei-Zhen HU
Elida Lailiya ISTIQOMAH
Yuhei ITO
Tomoko IWASHITA
David JENNENS
Lili JIANG
Morten Dam JOERGENSEN
Varchaswi KASHYAP
Petr KATRENKO
Shoaib KHALID
Wajid Ali KHAN
Geon-Bo KIM
Jinsu KIM
Jongkuk KIM
Kyungwon KIM
Tomoe KISHIMOTO
Kenji KIUCHI
Tomoyuki KONNO
Hirohisa KUBOTA
Ka Hei Martin KWOK
Jason LEE
Bo LI
Lan-Chun LV
Harvey Jonathan MADDOCKS
Hiroki MAKINO
Tsunayuki MATSUBARA
Satoru MATSUMOTO
Takuya MINAKUCHI
Junya NAKAMURA
Hiroshi NAKANO
Jeremy NEVEU
Shoichiro NISHIMURA
Abdullah NOORAIHAN
Hideyuki OIDE
Kou OISHI
Elisabeth PANZENBOECK
Nikhul PATEL
Lemuel, Jr. PELAGIO
Almut Maria PINGEL
Michael PRIM
Camila RANGEL SMITH
Kanishka RAWAT
Rintarn SAENGSAI
Saurabh SANDILYA
Diego SEMMLER
Mehar SHAH
Manoj Kumar SINGH
Yuji SUDO
Kento SUZUKI
Kong Guan TAN
Jia Jian TEOH
Ma TIAN
Nam TRAN
Po-Yen TSENG
Cenk TÜRKOGLU
Annika VAUTH
Vitaly VOROBYEV
Jun WAKABAYASHI
Sebastian WANDERNOTH
Su-Yin WANG
Ian WATSON
Nadeesha WICKRAMAGE
Peter WIJERATNE
Thomas WILLIAMS
Zhong-Zhi XIANYU
Benda XU
Daisuke YAMATO
Katsuya YAMAUCHI
Yung-Shun YEH
Ali ZAMAN
Guang ZHAO
Xiaokang ZHOU

Posters

Author	Poster title
S. CHAKRABORTY, R. ADHIKARI, A. DASGUPTA, S. ROY	Non standard Interaction in Neutrino Oscillations and the recent T2K and Daya-Bay Experiment
C. COUTURIER	Tests of Lorentz invariance with γ-ray observatories
D. YAMATO, T. OKUSAWA, Y. SEIYA, K. YAMAMOTO (ON BEHALF OF THE CDF COLLABORATION)	Search for the Higgs Boson Produced in Association with a Vector Boson Using Like-Sign Dilepton Events at CDF
F. DORDEI	Lifetime measurements for exclusive b$\rightarrow J/\Psi\xi$ decays with $J/\Psi \rightarrow \mu^+\mu^-$ at LHCb
E. CLEMENT	Search for heavy resonances decaying to long-lived neutral particles in leptonic channels
M. FRANCHINI	$t\bar{t}$ differential cross section in the lepton+jets channels at ATLAS and boosted object selection
M. HASHEMINIA, M.E. ZOMORRODIAN, A. MIRJALILI	The influence of fragmentation models in production of hadron jets in electron-positron annihilation
S. HIROSE	Performance Study of MCP-PMTs for the TOP in Belle II
V.K.S. KASHYAP, C. YADAV, S.T. SEHGAL, R. SEHGAL, R.G. THOMAS, L.M. PANT	Resistive Plate Chambers for INO and CMS
K. SUZUKI	Performance of the Muon Monitor in the T2K Experiment
K. KIM, ON BEHALF OF THE KIMS COLLABORATION	Studies of the Scintillator and Reflective foils in the KIMS detector

Author	Poster title
T. KISHIMOTO	Search for the SM Higgs boson in H→WW→ $l\nu l\nu$ with ATLAS
K.G. TAN ON BEHALF OF THE ATLAS COLLABORATION	Tau Reconstruction and Identification Algorithms and Performance at the ATLAS Experiment
M. PRIM	Angular analysis and search for CP-violation in B→ $\phi(K\pi)^*_0$ at Belle
H. NAKANO, A. ISHIKAWA, K. SUMISAWA, H. YAMAMOTO	Optimization of B0 to Ks eta gamma decay mode reconstruction for time-dependent CPV search
S. NISHIMURA	Precise muon g-2 measurement and searching for muon EDM at J-PARC
K. RAWAT, V. BHATNAGAR, D. INDUMATHI	Study of muon resolution in the INO-ICAL Detector
S. SANDILYA, G. MOHANTY, K. TRABELSI	Exclusive search for η_b(1S) and η_b(2S) in radiative Υ(2S) decays at Belle
S. WANDERNOTH	Measurement of Δm_s in the decay $B^0_s \to D^-_s \pi^+$ at LHCb
N. TRAN	A Search for μ – e Conversion and a Study of Muon Capture Backgrounds
A. VAUTH	Design of a Quartz Cherenkov Detector for Polarimetry at the ILC