CERN–2006–015
19 December 2006

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

2005 CERN–CLAF School of High-Energy Physics

Malargüe, Argentina
27 February–12 March 2005

Proceedings

Editors: N. Ellis
M.T. Dova

GENEVA
2006

CERN–290 copies printed–December 2006

Abstract

The CERN–CLAF 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 field theory and the Standard Model, quantum chromodynamics, CP violation and flavour physics, as well as reports on cosmic rays, the Pierre Auger Project, instrumentation, and trigger and data-acquisition systems.

Preface

The third in the new series of Latin American Schools of High-Energy Physics took place in Malargüe, located in the south-east of the Province of Mendoza in Argentina, from 27 February to 12 March 2005. It was organized jointly by CERN and CLAF (Centro Latino Americano de Física), and with the strong support of CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas). Fifty-four students coming from eleven different countries attended the School. While most of the students stayed in Hotel Rio Grande, a few students and the Staff stayed at Microtel situated close by. However, all the participants ate their meals together at Hotel Rio Grande. According to the tradition of the School the students shared twin rooms mixing nationalities and in particular Europeans together with Latin Americans.

María Teresa Dova from La Plata University was the local director for the School. The lectures were given at the Expositions and Convention Centre, 'Thesaurus', an ultra-modern building, some three hundred metres from the two hotels and just across the road from the Pierre Auger Observatory central campus. The proximity of the Auger Observatory was the main reason for choosing Malargüe as the site for the School, and a special presentation on the Auger experiment, given by Alan Watson the present project spokesman, was highly appreciated by the school participants. The presentation was followed by an interesting tour of Los Leones where one of the Auger fluorescence detectors is installed and to some of the detectors of the surface array deployed in Pampa Amarilla.

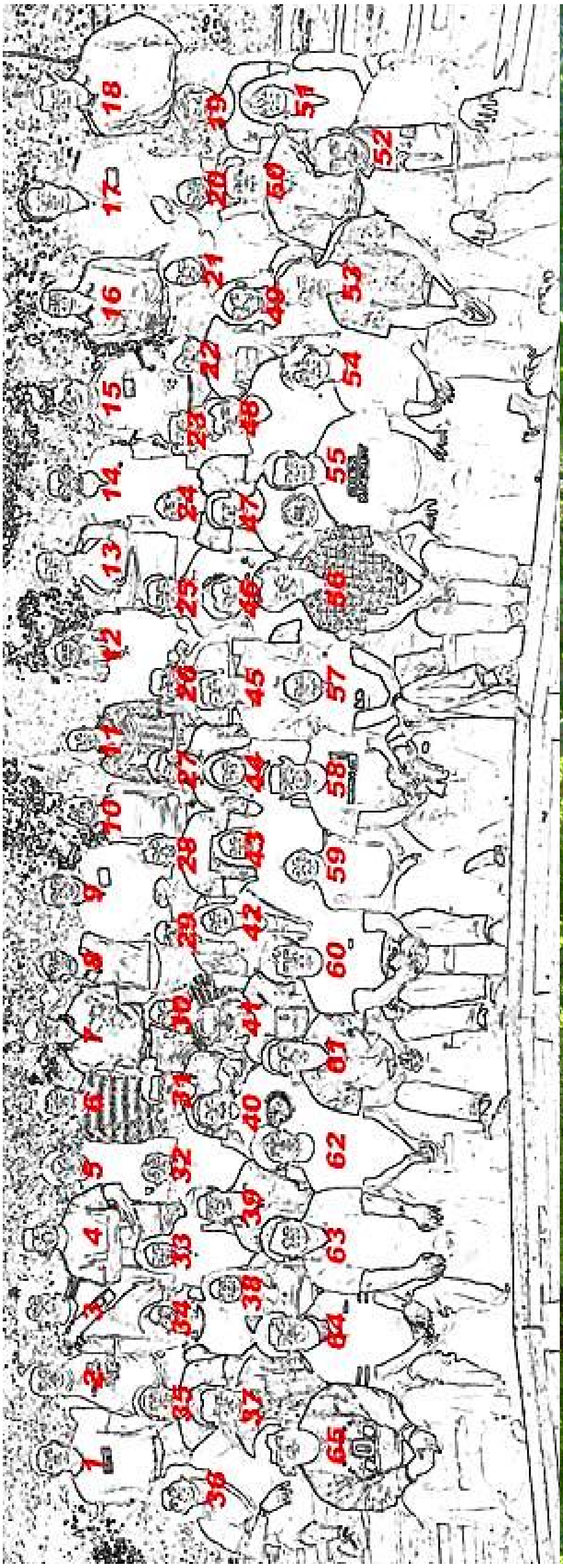
Our sincere thanks go to Tere who, together with the local committee, made it possible to organize the School and contributed to its success, and to Esteban Roulet, Ricardo Piegaia and Juan Tirao, who helped during the visit to the different locations which had been proposed for the School. We are also grateful to CONICET for their financial support. Our thanks are due to the lecturers and discussion leaders for their active participation in the School and for making the scientific programme so stimulating. The students, who in turn manifested their good spirits during two intense weeks, undoubtedly appreciated the personal contributions of the teaching staff in answering questions and explaining points of theory.

We are very grateful to Danielle Métral for her untiring efforts in the lengthy preparations for and the day-to-day care of the School. Her continuous care of the participants and their needs was highly appreciated. We are also grateful to Andrea Aparicio for her friendly assistance with translations, accounting, and other practical problems. Our special thanks also go to Graciella Viollaz, head of the local Tourist Office, Raul Rodriguez, Mayor of Malargüe, and Celso A. Jaque, former Mayor and now National Senator, who invited the school to Malargüe, providing the local infrastructure for free.

The School participants enjoyed two memorable excursions, one to Castillos de Pincheira, a natural and impressive geologic structure about thirty kilometres from the city of Malargüe, and another to Las Leñas, Argentina's largest ski resort. On the way to Las Leñas stops were made at the Well of the Souls ('Pozo de las Ánimas'), and the lagoon of the Enchanted Girl ('La Niña Encantada').

However, the success of the School was to a large extent due to the students themselves. Their posters were of excellent quality both technically and in content, and throughout the School they participated actively during the lectures, in the discussion sessions, and with genuine interest in the different activities and excursions.

Egil Lillestøl
on behalf of the Organizing Committee


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

CHICK MAGNET

People in the photograph

1	Sánchez Paredes, Marcial
2	Cobos-Martínez, Jesús
3	González-Canales, Félix Francisco
4	Schneebeli, Matthias
5	Cortes Maldonado, Ismael
6	Nunokawa, Hiroshi
7	Lillestøl, Egil
8	Stark Schneebeli, Luisa Sabrina
9	Asorey, Claudio
10	Goy López, Silvia
11	Paniccia, Mercedes
12	Olinto, Angela
13	Mello Jr., Walter
14	Racker, Juan
15	Capa Tira, Claudio
16	Mondragón, Miguel
17	Swain, John
18	Contreras, Guillermo
19	Métral, Danielle
20	Paulucci, Laura
21	Marinho, Franciole
22	Castromonte Flores, César Manuel
23	Almeida, Rogerio
24	Rangel, Murilo
25	Busca, Nicolas
26	Pioppi, Michele
27	Gómez Dumm, Daniel
28	Moreno, Juan Cruz
29	Cabarcas, Jose Miguel
30	Schoch Vianna, Cristina
31	Moreno Briceño, Alexander
32	Akiba, Kazuyoshi
33	Morales Morales, Cristina
34	Fry, Catherine
35	Sepúlveda, Fernando
36	Simili, Emanuele
37	Kharzeev, Dmitri
38	Uranga, Llinersy
39	Bazo Alba, José Luis
40	Castillo, Felisola Óscar Alberto
41	Ellis, John
42	Blanco Siffert, Beatriz
43	Deplano, Caterina
44	Dabrowski, Anne
45	Aliaga Soplín, Leonidas
46	Vázquez-Jauregui, Eric
47	Tamashiro, Alejandro
48	Pérez, Ricardo
49	Martínez, Santiago Andrés
50	Sommers, Paul
51	Dova, María Teresa
52	Fonseca, Sandro
53	Ellis, Nick
54	Medina, María Clementina
55	García, Carlos
56	Watson, Alan
57	Ramírez Rojas, Alba
58	Morales, Allan
59	Ruiz Tabasco, Julia Elizabeth
60	González Sprinberg, Gabriel
61	Peinado-Rodríguez, Eduardo
62	Helo Herrera, Juan Carlos
63	Barrett, Matthew
64	Blanco Covarrubias, Ernesto Alejandro
65	Bravo, Alberto

Contents

Introductory lectures on quantum field theory[*]

L. Álvarez-Gaumé [a] and M.A. Vázquez-Mozo [b, c]
[a] CERN, Geneva, Switzerland
[b] Departamento de Física Fundamental, Universidad de Salamanca, Salamanca, Spain
[c] Instituto Universitario de Física Fundamental y Matemáticas (IUFFyM), 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

The audience for these lectures was composed to a large extent of 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 these lectures are intended to provide a review of those parts of the subject to be used later by other lecturers. Although a cursory acquaintance with the subject of quantum field theory is helpful, the only requirement necessary to follow the lectures is a working knowledge of quantum mechanics and special relativity.

The guiding principle in choosing the topics (apart from to serve as introductions to later courses) was to cover 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 following.

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

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

Because of 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. Full details and proofs can be found in the many textbooks on the subject, and in particular in the ones provided in the list of references [1–9]. Especially modern presentations, very much in the spirit of these lectures, can be found in Refs. [4, 5, 9]. We should nevertheless warn the reader that we have been a little cavalier about references. Our aim has been to provide mostly a (non-exhaustive) list of references for further reading. We apologize to those authors who feel misrepresented.

[*] Based on lectures delivered by L. Álvarez-Gaumé.

1.1 A note about notation

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

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

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 make use of 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 already clear 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 waves 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

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

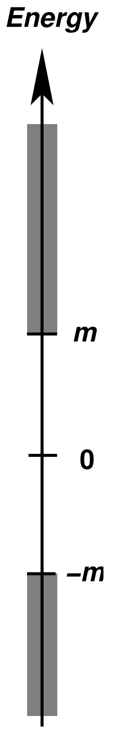


Fig. 1: Spectrum of the Klein–Gordon wave equation

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:

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

The wave functions $f_p(t,x)$ describe states with momentum $\vec{p}$ and energy given by ω_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:

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

Therefore the energy spectrum of the theory satisfies $|E|>m$ and is unbounded from below (see Fig. 1). Although in the 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

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

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

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

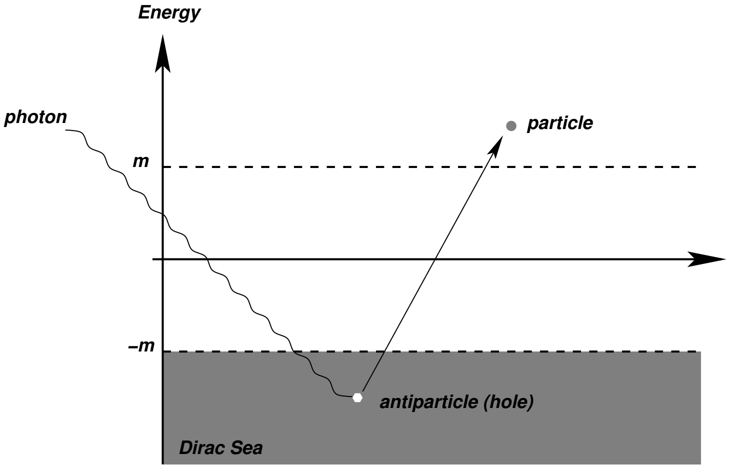


Fig. 2: Creation of a particle–antiparticle pair in the Dirac sea picture

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

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

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 from the conclusion that interactions will produce particle–antiparticle pairs 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 upon an apparent paradox when trying to describe the scattering of a relativistic electron by a square potential using Dirac's wave equation [10] (for pedagogical reviews see Refs. [11, 12]). In order to capture the essence of the problem without entering into unnecessary complication we shall 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 shown in Fig. 3. A solution to the wave equation in regions I and II is given by

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

where the mass-shell condition implies that

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

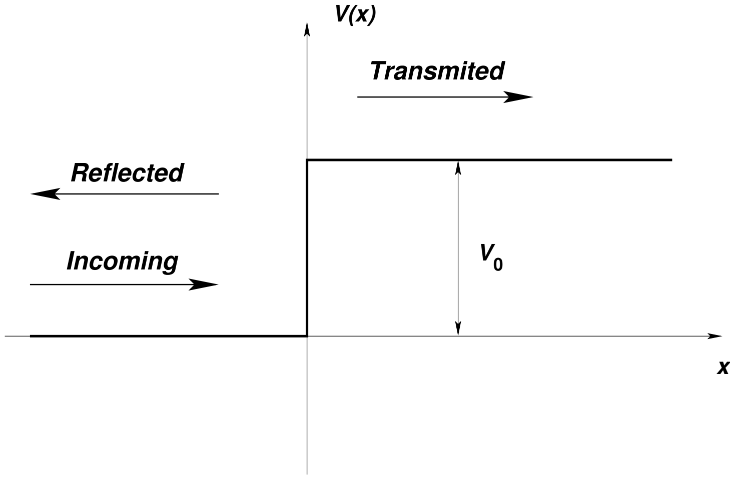


Fig. 3: Illustration of the Klein paradox

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 behaviour similar to that 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 expects to find only 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 $E - m < V_0$ *and* $E - m < V_0 - 2m$ 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 $V_0 - 2m < E - m < V_0$ 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 non-vanishing 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 on giving a single-particle interpretation to the relativistic wave function. Actually, a multiparticle analysis of the paradox [11] shows that what happens when $E - m > V_0 - 2m$ is that the reflection of the incoming particle by the barrier is accompanied by the creation of particle–antiparticle pairs out of the energy of the barrier. (Note that for this to happen it is required that $V_0 > 2m$, the threshold for the creation of a particle–antiparticle pair.)

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 distance scales larger than $1/m$. This case was worked out by Sauter shortly after Klein pointed out the paradox [13]. 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 $V(x) = \frac{V_0 x}{d}$. He found that when $d > \frac{1}{m}$ the transmission coefficient is exponentially small[1].

[1] In Section (8.1) we shall see how, in the case of the Dirac field, this exponential behaviour can be associated with the

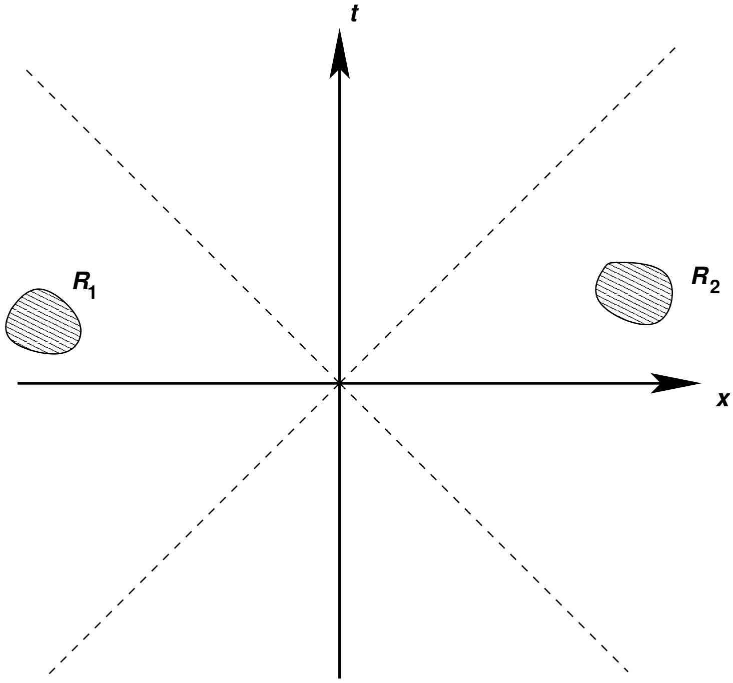


Fig. 4: Two regions R_1, R_2 that are causally disconnected

The creation of particles is impossible to avoid whenever one tries to locate a particle of mass m within its Compton wavelength. Indeed, from the 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 operators 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

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

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

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

creation of electron–positron pairs due to a constant electric field (Schwinger effect).

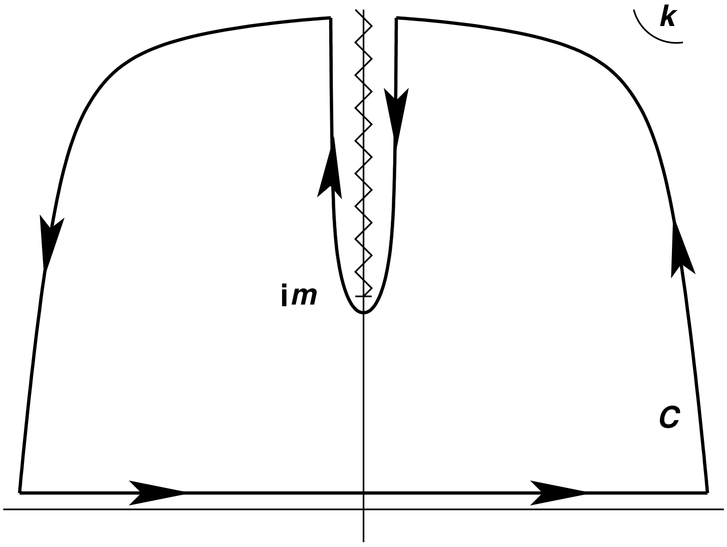


Fig. 5: Complex contour C for the computation of the integral in Eq. (24)

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,

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

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

Therefore, relativistic invariance forces the introduction of quantum fields. It is only when we insist on keeping a single-particle interpretation that we come up 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 on interpreting the wave function $\psi(t,\vec{x})$ as the probability density of finding the particle at location $\vec{x}$ in 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 shall see below.

In any scattering experiment, the only information available to us is the set of quantum numbers associated with the set of free particles in the initial and final states. Ignoring for the moment other quantum numbers like spin and flavour, 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

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

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 taken the integration measure to be 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 $a^\dagger(\vec{p})$,

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

Covariance under spatial rotations is all we need if we are interested in a non-relativistic theory. However, in a relativistic quantum field theory we must preserve more than $\mathrm{SO}(3)$; in fact we need the expressions to be covariant under the full Poincaré group $\mathrm{ISO}(1,3)$ consisting of 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 $\theta(x)$ represents 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 corresponds 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 $\widehat{P}^\mu$:

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

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 non-relativistic basis of states $\{|\vec{p}\rangle\}$ as

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

Using now Eq. (28) for the non-relativistic states, expression (39) follows provided

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

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

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 have done so far, we have succeeded 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 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 Eq. (47) and the fact that the measure is invariant under Lorentz transformation, one can easily show that $\widehat{P}^\mu$ transforms 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

$$\begin{aligned} [\alpha(\vec{p}),\alpha^\dagger(\vec{p}\,')] &= (2\pi)^3(2\omega_p)\delta(\vec{p}-\vec{p}\,')\;, \\ [\alpha(\vec{p}),\alpha(\vec{p}\,')] &= [\alpha^\dagger(\vec{p}),\alpha^\dagger(\vec{p}\,')] = 0\;. \end{aligned} \tag{51}$$

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 then be 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 an 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 these states are Lorentz invariant can be checked by noting 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 commutators of two fields have to vanish outside the relative light-cone

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

- **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 lead 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 also be computed to be

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

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

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

where $\varepsilon(x)$ is defined as

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

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. Note 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 of the quantum scalar field: existence of asymptotic states, locality, microcausality and relativistic invariance. With only these ingredients we have managed to go quite far. The above 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 role, it is important to check at the end of the calculation that the resulting quantum theory is Lorentz invariant. In the following we shall briefly overview the canonical quantization of the Klein–Gordon scalar field.

The starting point is the action function $S[\phi(x)]$ 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 brackets. Given two functions $A[\phi,\pi]$, $B[\phi,\pi]$ 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 at 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 has 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 an infinite collection of harmonic oscillators per unit volume, each one labelled by $\vec{p}$. Even if those oscillators are not excited, they contribute to the vacuum energy with their zero-point energy, given by $\frac{1}{2}\omega_p$. This vacuum contribution to the energy 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 discard this infinite contribution to the energy of the vacuum as unphysical, it has observable consequences. In 1948 Hendrik

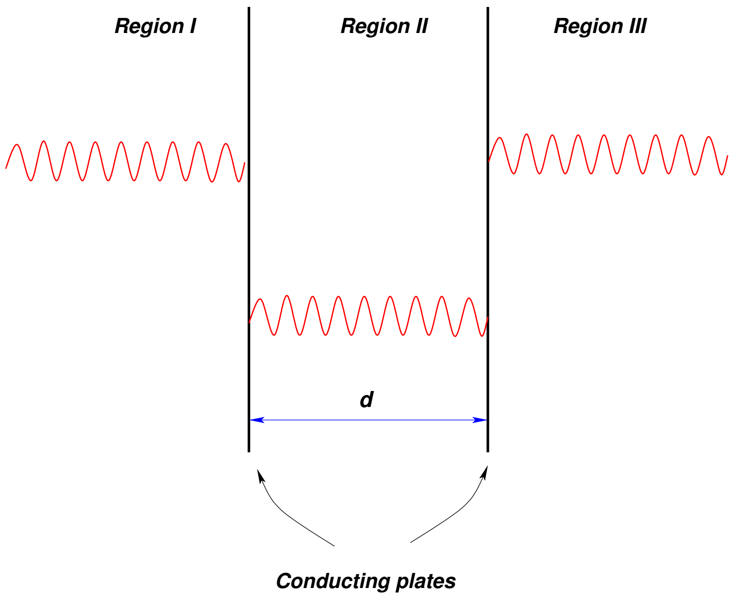


Fig. 6: Illustration of the Casimir effect. In Regions I and II the spectrum of modes of the momentum $p_\perp$ is continuous, while in the space between the plates (Region II) it is quantized in units of $\frac{\pi}{d}$

Casimir pointed out [14] that although a formally divergent vacuum energy would not be observable, any variation in this energy would be (see Ref. [15] for comprehensive reviews).

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

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

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

This subtraction cancels the contribution of the modes outside the plates. Because of the boundary conditions imposed by the plates, the momenta 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

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

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 cut-off. 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 [16]:

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

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 cut-off, $\Lambda \to \infty$, 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 [17] and since then has been checked with better and better precision in a variety of situations [15].

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 behaviour under Lorentz transformations. Before considering other fields we pause and study the properties of the Lorentz group.

[4] Actually, one could introduce any cut-off 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 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:

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

Using Eq. (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 $\mathrm{SU}(2) \approx$ 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 self-dual and an antiself-dual antisymmetric rank-two 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
$(\mathbf{\frac{1}{2}},\mathbf{0})$	Right-handed spinor
$(\mathbf{0},\mathbf{\frac{1}{2}})$	Left-handed spinor
$(\mathbf{\frac{1}{2}},\mathbf{\frac{1}{2}})$	Vector
$(\mathbf{1},\mathbf{0})$	Self-dual antisymmetric 2-tensor
$(\mathbf{0},\mathbf{1})$	Antiself-dual antisymmetric 2-tensor

To conclude our discussion of the representations of the Lorentz group, we note 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 $(\mathbf{\frac{1}{2}},\mathbf{\frac{1}{2}})$ is invariant under parity, whereas a left-handed Weyl spinor $(\mathbf{\frac{1}{2}},\mathbf{0})$ transforms into a right-handed one $(\mathbf{0},\mathbf{\frac{1}{2}})$ and vice versa.

4.2 Spinors

Weyl spinors

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

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

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:

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

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

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

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

we are left with just one possibility up to a sign:

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

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 on demanding that the Lagrangian be invariant under the global symmetry $u_\pm \longrightarrow e^{i\theta}u_\pm$ 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 correspond to the antisymmetric combinations

$$\epsilon_{ab}u^a_\pm u^b_\pm, \tag{108}$$

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 satisfy the Fermi–Dirac statistics and therefore satisfy anticommutation relations, whereas fields of integer spin follow the statistics 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:

$$u^a_\pm u^b_\pm + u^b_\pm u^a_\pm = 0\,. \tag{109}$$

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 naive definition of the classical limit as a limit of large occupation numbers cannot be applied. Fermion fields do not really make sense classically.

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

$$\mathcal{L}^{\pm}_{\text{Weyl}} = iu^{\dagger}_{\pm}\sigma^{\mu}_{\pm}\partial_{\mu}u_{\pm} + \frac{1}{2}m\epsilon_{ab}u^{a}_{\pm}u^{b}_{\pm} + \text{h.c.} \tag{110}$$

This mass term, said to be 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, for 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 problem if we introduce Majorana masses for right-handed neutrinos, since they are singlets under all Standard Model gauge groups. Such a term will, however, break the global U(1) lepton number charge because the operator $\epsilon_{ab}\nu^{a}_{R}\nu^{b}_{R}$ 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 of Weyl fermions u_{+} and u_{+}. Actually, these two fields can be combined in a single four-component spinor

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

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

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

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

These equations of motion can be derived from the Lagrangian density

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

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

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

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 Eq. (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 A) 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, $k^2=m^2$.

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

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

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 $\widehat{b}^\dagger_\alpha(\vec{k},s)$, $\widehat{b}_\alpha(\vec{k})$ create and annihilate, respectively, 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]

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

In the case of $d_a(\vec{k},s)$, $d^\dagger_a(\vec{k},s)$ we have a set of creation–annihilation operators for the corresponding antiparticles (for example, positrons). This is clear if we notice that $d^\dagger_a(\vec{k},s)$ can be seen as the annihilation operator of a negative energy state of the Dirac equation with wave function $v_a(\vec{k},s)$. 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

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

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

The Hamiltonian operator for the Dirac field is

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

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

[5] To simplify notation, and since there is no risk of confusion, from now on we drop the hat ($\widehat{\ }$) when indicating operators.

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

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

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 8.2 we will see that in a certain type of theory 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

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

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 shall explain further in Section 5, this current is associated with the invariance of the Dirac Lagrangian under the global phase shift $\psi \longrightarrow e^{i\theta}\psi$. 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

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

Classically $(\varphi, \vec{A})$ are seen as only a convenient way to solve Maxwell's 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 $j^\mu = (\rho, \vec{\jmath})$ 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 role 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 gives rise to surprising phenomena. The first evidence of the role played by the electromagnetic potentials at the quantum level was pointed out by Yakir Aharonov and David Bohm [18]. 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 non-vanishing 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 shall observe the

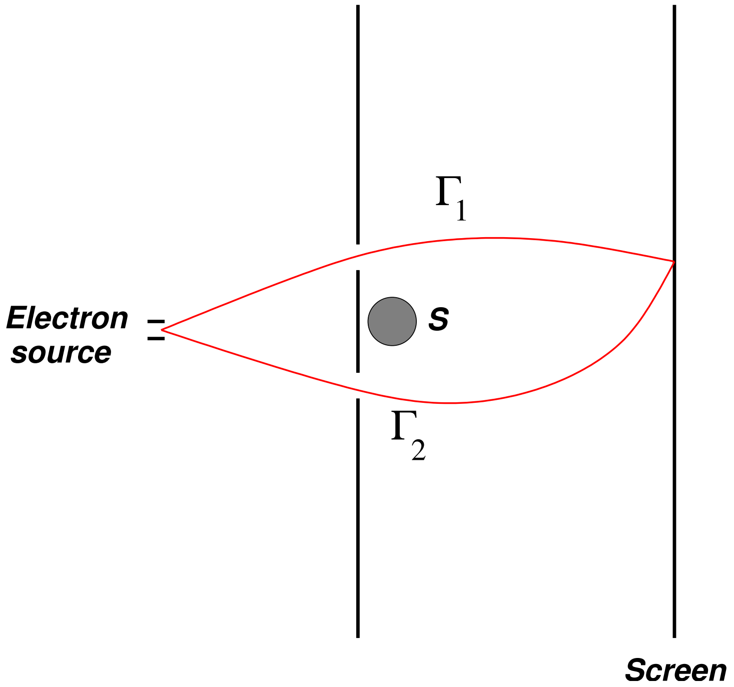


Fig. 7: Illustration of an interference experiment to show the Aharonov–Bohm effect. S represents the solenoid within which the magnetic field is confined.

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

The first question that comes up is what happens with gauge invariance. Since we said that $\vec{A}$ can be changed by a gauge transformation it seems that the resulting interference patterns might depend on the gauge used. Actually, the phase U in Eq. (145) is independent of the gauge although, unlike other gauge-invariant quantities like $\vec{E}$ and $\vec{B}$, is non-local. Note 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 $\theta = \pi/2$ 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 $(\rho_m, \vec{\jmath}_m)$ in such a way that the duality (146) is restored when supplemented by the transformation

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

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

In 1931 Dirac [19] studied the possibility of finding solutions of the completed Maxwell equation with a magnetic monopole of charge g, i.e., solutions to

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

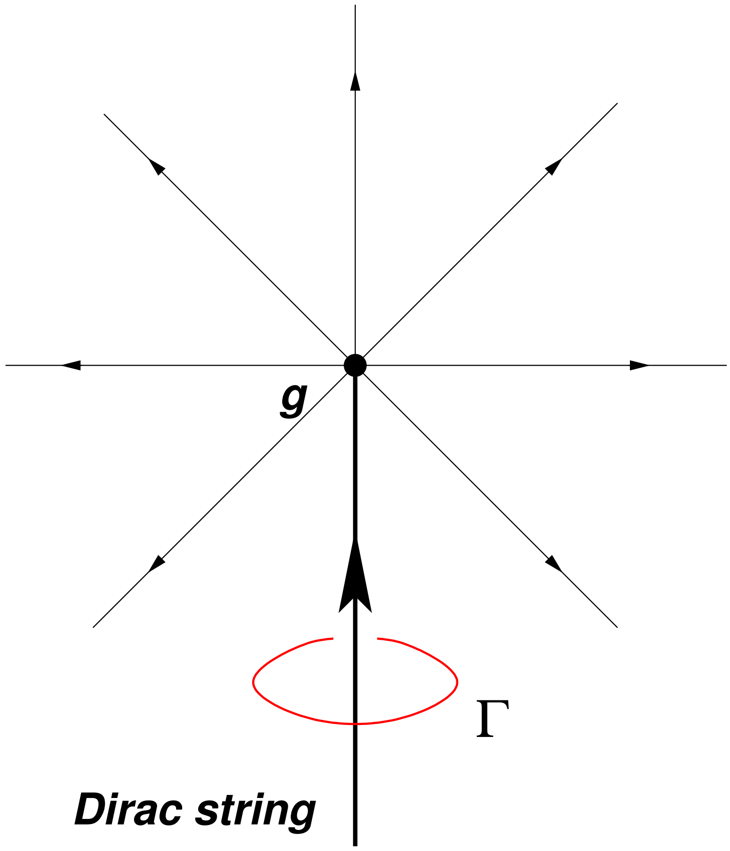


Fig. 8: The Dirac monopole

Away from the position of the monopole, $\vec{\nabla}\cdot\vec{B}=0$ and the magnetic field can still be derived locally from a vector potential $\vec{A}$ according to $\vec{B}=\vec{\nabla}\times\vec{A}$. However, the vector potential cannot be regular everywhere since otherwise Gauss's law would imply that the magnetic flux threading a closed surface around the monopole should vanish, contradicting Eq. (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; 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 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 long 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 picks 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 monopole somewhere in the Universe implies for consistency the quantization of the electric charge in units of $\frac{2\pi}{g}$, where g is 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 $\rho = 0$, $\vec{\jmath} = \vec{0}$. 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 fewer 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

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

Note 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:

$$\widehat{A}_\mu(t,\vec{x}) = \sum_{\lambda=\pm1}\int\frac{d^3k}{(2\pi)^3}\frac{1}{2|\vec{k}|}\left[\epsilon_\mu(\vec{k},\lambda)\widehat{a}(\vec{k},\lambda)e^{-i|\vec{k}|t+i\vec{k}\cdot\vec{x}} + \epsilon_\mu(\vec{k},\lambda)^*\,\widehat{a}^\dagger(\vec{k},\lambda)e^{i|\vec{k}|t-i\vec{k}\cdot\vec{x}}\right] \tag{155}$$

where $\lambda = \pm1$ represent the helicity of the photon, and $\epsilon_\mu(\vec{k},\lambda)$ are solutions to the equations of motion with well-defined momentum and helicity. Because of Eq. (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

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

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 to gauge invariance. In particular the gauge freedom seems to introduce states in the Hilbert space with negative probability. A careful analysis shows that when gauge invariance is properly handled these spurious states decouple from physical states and can be eliminated. The details can be found in standard textbooks [1–9].

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 for by a U(1) phase shift in the wave function. In the case of the Dirac equation we know that the Lagrangian is invariant under $\psi \rightarrow e^{ie\varepsilon}\psi$, with 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, $\psi \rightarrow e^{ie\varepsilon(x)}\psi$, it suffices to replace the ordinary derivative ∂_μ with 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) describes an interacting theory. By plugging Eq. (159) into the Lagrangian we find 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 is plagued with infinities that should be taken care of. We shall return to this problem in Section 7.

Non-Abelian gauge theories

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

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

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

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 transformation 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

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

Note that this definition of $F_{\mu\nu}^a$ 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 non-Abelian 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

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

With the help of this we can write a generic Lagrangian for a non-Abelian 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:

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

[6] Some basic facts about Lie groups are summarized in Appendix A.

Most of the time 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 [20, 21].

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 does 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 Ref. [20] we use the symbol $\approx$ to indicate a 'weak' equality that holds when the constraints $f_a(q,p) = 0$ are satisfied. Note, however, that since the computation of the Poisson brackets involves derivatives, the constraints can be used only after the bracket has been 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 does 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 '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 [20]. 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

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

[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 is the case with the Lagrangian $L(q,\dot{q}) = q$.

we find that in 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

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

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's law $\vec{\nabla}\cdot\vec{E} = 0$. These two constraints have vanishing Poisson brackets 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 by taking $\lambda_1 = 0$. Note 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's 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's 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. Note that the constraint $\vec{\nabla}\cdot\vec{E} = 0$ can be implemented by demanding $\vec{\nabla}\cdot\vec{A} = 0$, which reduces the three degrees of freedom of $\vec{A}$ to the two physical degrees of freedom of the photon.

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

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

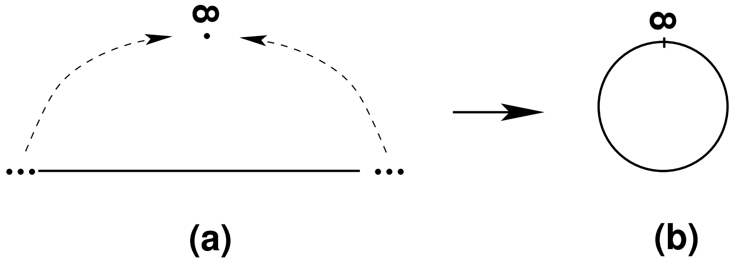


Fig. 9: Compactification of the real line (a) into the circumference S^1 (b) by adding the point at infinity

By definition, physical states should not change when a gauge transformation is performed. This is implemented by requiring that the operator $\mathcal{U}(\varepsilon)$ act 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's 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 role of gauge invariance [22]. As we have learned, the existence of a gauge symmetry in a theory reflects a degree of redundancy in the description of physical states in terms of the degrees of freedom appearing in the Lagrangian. In classical mechanics, for example, the state of a system is usually determined by the value of the canonical coordinates (q_i, p_i). We know, however, that this is not the case for constrained Hamiltonian systems where the transformations generated by the first-class constraints change the value of q_i and p_i without changing the physical state. In the case of Maxwell theory, for every physical configuration determined by the gauge invariant quantities $\vec{E}$, $\vec{B}$ there is an infinite number of possible values of the vector potential that are related by gauge transformations $\delta A_\mu = \partial_\mu\varepsilon$.

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 $A_0 = 0$, $\vec{\nabla}\cdot\vec{A} = 0$ select 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 role. To illustrate the point let us first deal with a toy model of a U(1) gauge theory in 1+1 dimensions. Later we shall 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 non-trivial 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

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

and the conditions $g(x)^\dagger g(x) = \mathbf{1}$, $\det g = 1$ imply 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 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 analysed we find that because of the non-trivial 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's law annihilates physical states. For a non-Abelian 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 here is that only the elements of $\mathcal{G}_0$ can be written as exponentials of the infinitesimal generators. Since these 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's 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 $n = 1$

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

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

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

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

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

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

To find a physical interpretation of this result we are going to look for similar things in other physical situations. One of them is borrowed from condensed-matter physics and refers to the quantum states of electrons in the periodic potential produced by the ion lattice in a solid. For simplicity we discuss the one-dimensional case where the minima of the potential are separated by a distance a. When the barrier between consecutive degenerate vacua is high enough we can neglect tunnelling 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 non-Abelian gauge theories. The vacuum state labelled by θ plays a role 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 non-Abelian 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^i_a and we assume that the Gauss's 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

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

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 non-trivial topology of the gauge group has 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 non-Abelian version of Maxwell theory we consider only 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

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

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 Eq. (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 non-Abelian 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 non-vanishing electric dipole moment for the neutron [23]. The fact that this is not observed imposes 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 also be obtained using path integral techniques from the Lagrangian (207). The second term in Eq. (207) gives then a contribution that depends on the winding number of the corresponding gauge configuration.

5 Symmetries

5.1 Noether's theorem

In classical mechanics and classical field theory there is a basic result that relates symmetries and conserved charges. This is called Noether's theorem and states that for each continuous symmetry of the system there is conserved current. In its simplest version in classical mechanics it can be easily proved.

Let us consider a Lagrangian $L(q_i, \dot{q}_i)$ which is invariant under a transformation $q_i(t) \rightarrow q_i'(t, \epsilon)$ labelled by a parameter ϵ. This means that $L(q', \dot{q}') = L(q, \dot{q})$ without using the equations of motion[9]. 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

$$0 = \delta_\epsilon L(q_i, \dot{q}_i) = \frac{\partial L}{\partial q_i}\delta_\epsilon q_i + \frac{\partial L}{\partial \dot{q}_i}\delta_\epsilon \dot{q}_i = \left[\frac{\partial L}{\partial q_i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}\right]\delta_\epsilon q_i + \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_i}\delta_\epsilon q_i\right). \tag{210}$$

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

$$\dot{Q} = 0 \quad \text{with} \quad Q \equiv \frac{\partial L}{\partial \dot{q}_i}\delta_\epsilon q_i \, . \tag{211}$$

Note that in this derivation it is crucial that the symmetry depend on a continuous parameter since otherwise the infinitesimal variation of the Lagrangian in Eq. (210) 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

$$\delta_\epsilon \mathcal{L} = \partial_\mu K^\mu \, . \tag{212}$$

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

$$\partial_\mu K^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)}\partial_\mu \delta_\epsilon \phi + \frac{\partial \mathcal{L}}{\partial \phi}\delta_\epsilon \phi = \partial_\mu\left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)}\delta_\epsilon \phi\right) + \left[\frac{\partial \mathcal{L}}{\partial \phi} - \partial_\mu\left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)}\right)\right]\delta_\epsilon \phi \, . \tag{213}$$

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

$$\partial_\mu J^\mu = 0 \quad \text{with} \quad J^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)}\delta_\epsilon \phi - K^\mu \, . \tag{214}$$

Actually a conserved current implies the existence of a charge

$$Q \equiv \int d^3x \, J^0(t, \vec{x}) \tag{215}$$

which is conserved

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \int d^3x \, \partial_0 J^0(t, \vec{x}) = -\int d^3x \, \partial_i J^i(t, \vec{x}) = 0 \, , \tag{216}$$

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. (215) is promoted to an operator that generates the symmetry on the fields

$$\delta\phi = i[\phi, Q] \, . \tag{217}$$

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

$$\phi'(x) - \phi(x) = -a^\mu \partial_\mu \phi + \mathcal{O}(a^2) \quad \Longrightarrow \quad \delta\phi = -a^\mu \partial_\mu \phi \, . \tag{218}$$

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

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

$$\delta\mathcal{L} = -a^{\mu}\partial_{\mu}\mathcal{L} \, . \tag{219}$$

Therefore the corresponding conserved charge is

$$J^{\mu} = -\frac{\partial\mathcal{L}}{\partial(\partial_{\mu}\phi)}a^{\nu}\partial_{\nu}\phi + a^{\mu}\mathcal{L} \equiv -a_{\nu}T^{\mu\nu} \, , \tag{220}$$

where we introduced the energy–momentum tensor

$$T^{\mu\nu} = \frac{\partial\mathcal{L}}{\partial(\partial_{\mu}\phi)}\partial^{\nu}\phi - \eta^{\mu\nu}\mathcal{L} \, . \tag{221}$$

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

$$\partial_{\mu}T^{\mu\nu} = 0 \, . \tag{222}$$

The associated conserved charges are given by

$$P^{\nu} = \int d^3x \, T^{0\nu} \tag{223}$$

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 to a physically relevant conserved charge is the global phase invariance of the Dirac Lagrangian (117), $\psi \longrightarrow e^{i\theta}\psi$. 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

$$j^{\mu} = \overline{\psi}\gamma^{\mu}\psi \, , \qquad \partial_{\mu}j^{\mu} = 0 \, , \tag{224}$$

thus implying the existence of a conserved charge

$$Q = \int d^3x\overline{\psi}\gamma^0\psi = \int d^3x\psi^{\dagger}\psi \, . \tag{225}$$

In physics there are several instances of global U(1) symmetries that act as phase shifts on spinors. This is the case, for example, for baryon and lepton number conservation in the Standard Model. A more familiar case is the U(1) local symmetry associated with electromagnetism. Note that although in this case we are dealing with a local symmetry, $\theta \rightarrow e\alpha(x)$, the invariance of the Lagrangian holds in particular for global transformations and therefore there is a conserved current $j^{\mu} = e\overline{\psi}\gamma^{\mu}\psi$. 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 non-Abelian unitary global symmetries acting as

$$\psi_i \longrightarrow U_{ij}\psi_j \, , \qquad U^{\dagger}U = \mathbf{1} \tag{226}$$

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

$$U = \exp\left(i\alpha_a T^a\right) \, , \qquad (T^a)^{\dagger} = T^a \, , \tag{227}$$

we find the conserved current

$$j^{\mu\, a} = \overline{\psi}_i T^a_{ij}\gamma^\mu\psi_j, \qquad \partial_\mu j^\mu = 0\ . \tag{228}$$

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

$$\begin{pmatrix} u \\ d \end{pmatrix} \longrightarrow M \begin{pmatrix} u \\ d \end{pmatrix}\ , \qquad M \in \mathrm{SU(2)}\ . \tag{229}$$

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

5.2 Symmetries in the quantum theory

We have seen that in canonical quantization the conserved charges Q^a associated with 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:

$$[Q^a, H] = 0\ . \tag{230}$$

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

$$\mathcal{U}(\alpha)|0\rangle \equiv e^{i\alpha_a Q^a}|0\rangle = |0\rangle \qquad \Longrightarrow \qquad Q^a|0\rangle = 0\ . \tag{231}$$

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. (217) it is easy to prove that

$$\mathcal{U}(\alpha)\phi_i\mathcal{U}(\alpha)^{-1} = U_{ij}(\alpha)\phi_j\ , \tag{232}$$

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,

$$|i\rangle = \phi_i|0\rangle\ , \tag{233}$$

we find that because of the invariance of the vacuum (231) the states $|i\rangle$ transform in the same representation as ϕ_i:

$$\mathcal{U}(\alpha)|i\rangle = \mathcal{U}(\alpha)\phi_i\mathcal{U}(\alpha)^{-1}\mathcal{U}(\alpha)|0\rangle = U_{ij}(\alpha)\phi_j|0\rangle = U_{ij}(\alpha)|j\rangle\ . \tag{234}$$

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

$$e^{i\alpha_a Q^a}|0\rangle \neq |0\rangle \qquad \Longrightarrow \qquad Q^a|0\rangle \neq 0\,. \tag{235}$$

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

To illustrate the consequences of Eq. (235) we consider the example of a number of scalar fields φ^i ($i = 1, \ldots, N$) whose dynamics is governed by the Lagrangian

$$\mathcal{L} = \frac{1}{2}\partial_\mu\varphi^i\partial^\mu\varphi^i - V(\varphi)\,, \tag{236}$$

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

$$\delta\varphi^i = \epsilon^a (T^a)^i_j \varphi^j\,, \tag{237}$$

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

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

$$H = \int d^3x \left[\frac{1}{2}\pi^i\pi^i + \frac{1}{2}\vec{\nabla}\varphi^i \cdot \vec{\nabla}\varphi^i + V(\varphi)\right] \tag{238}$$

and look for the minimum of

$$\mathcal{V}(\varphi) = \int d^3x \left[\frac{1}{2}\vec{\nabla}\varphi^i \cdot \vec{\nabla}\varphi^i + V(\varphi)\right]\,. \tag{239}$$

Since we are interested in finding constant field configurations, $\vec{\nabla}\varphi = \vec{0}$ to preserve translational invariance, the vacua of the potential $\mathcal{V}(\varphi)$ coincide with the vacua of $V(\varphi)$. Therefore the minima of the potential correspond to the vacuum expectation values[10]:

$$\langle\varphi^i\rangle: \qquad V(\langle\varphi^i\rangle) = 0, \qquad \left.\frac{\partial V}{\partial\varphi^i}\right|_{\varphi^i = \langle\varphi^i\rangle} = 0\,. \tag{240}$$

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

$$(H^\alpha)^i_j\langle\varphi^j\rangle = 0\,. \tag{241}$$

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*. Note 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:

$$(K^A)^i_j\langle\varphi^j\rangle \neq 0\,. \tag{242}$$

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

These are 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

$$V(\varphi) = \left.\frac{\partial^2 V}{\partial\varphi^i\partial\varphi^j}\right|_{\varphi=\langle\varphi\rangle} (\varphi^i - \langle\varphi^i\rangle)(\varphi^j - \langle\varphi^j\rangle) + \mathcal{O}\left[(\varphi - \langle\varphi\rangle)^3\right] \tag{243}$$

and the mass matrix is

$$M^2_{ij} \equiv \left.\frac{\partial^2 V}{\partial\varphi^i\partial\varphi^j}\right|_{\varphi=\langle\varphi\rangle} . \tag{244}$$

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

$$\delta V(\varphi) = \epsilon^a \frac{\partial V}{\partial\varphi^i}(T^a)^i_j\varphi^j = 0 . \tag{245}$$

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

$$\frac{\partial^2 V}{\partial\varphi^i\partial\varphi^k}(T^a)^i_j\varphi^j + \frac{\partial V}{\partial\varphi^i}(T^a)^i_k = 0 . \tag{246}$$

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

$$M^2_{ik}(T^a)^i_j\langle\varphi^j\rangle = 0 . \tag{247}$$

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

$$M^2_{ik}(K^A)^i_j\langle\varphi^j\rangle = 0 . \tag{248}$$

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

$$V(\vec{\varphi}) = \frac{\lambda}{4}\left(\vec{\varphi}^{\,2} - a^2\right)^2 . \tag{249}$$

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

$$\langle\vec{\varphi}\rangle : \qquad \mathrm{SO}(3) \longrightarrow \mathrm{SO}(2)\ . \tag{250}$$

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 these 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 leaves invariant only 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 loss of symmetry. This is, however, not the case. The full quantum theory is symmetric under the whole symmetry group SO(3). This is reflected in the fact that the physical properties of the theory do not depend on the particular point of the sphere $\langle\vec{\varphi}\rangle\ ^2 = a^2$ that we have chosen. Different vacua are related by the full SO(3) symmetry and therefore should give the same physics.

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

$$\begin{pmatrix} u_{L,R} \\ d_{L,R} \end{pmatrix} \longrightarrow M_{L,R} \begin{pmatrix} u_{L,R} \\ d_{L,R} \end{pmatrix}, \qquad M_{L,R} \in \mathrm{SU}(2)_{L,R}\ . \tag{251}$$

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. Note, 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 non-zero 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 also appears in many places in condensed matter [24]. 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 associated Goldstone bosons 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 $\mathrm{SU}(2)\times\mathrm{U}(1)$

but, in order to reproduce Fermi's four-fermion theory of the β-decay, it was necessary that two of the vector fields involved be massive. Also in condensed-matter physics massive vector fields are required to describe certain systems, most notably in superconductivity.

The way out of 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

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \overline{D_\mu\Phi}D^\mu\Phi - \frac{\lambda}{4}\left(\overline{\Phi}\Phi - \mu^2\right)^2 \ , \tag{252}$$

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

$$\Phi \longrightarrow e^{i\alpha(x)}\Phi \ , \qquad A_\mu \longrightarrow A_\mu + \partial_\mu\alpha(x) \ . \tag{253}$$

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

$$\langle\Phi\rangle = \mu e^{i\vartheta_0} \longrightarrow \mu e^{i\vartheta_0 + i\alpha(x)} \tag{254}$$

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:

$$\Phi(x) = \left[\mu + \frac{1}{\sqrt{2}}\sigma(x)\right] e^{i\vartheta(x)} \ . \tag{255}$$

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 (253). This means that performing a gauge transformation with parameter $\alpha(x) = -\vartheta(x)$ we can get rid of the phase in Eq. (255). Substituting then $\Phi(x) = \mu + \frac{1}{\sqrt{2}}\sigma(x)$ in the Lagrangian we find

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + e^2\mu^2 A_\mu A^\mu + \frac{1}{2}\partial_\mu\sigma\partial^\mu\sigma - \frac{1}{2}\lambda\mu^2\sigma^2 \\ &- \lambda\mu\sigma^3 - \frac{\lambda}{4}\sigma^4 + e^2\mu A_\mu A^\mu\sigma + e^2 A_\mu A^\mu\sigma^2 \ . \end{aligned} \tag{256}$$

What is 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

$$m_\gamma^2 = 2e^2\mu^2 \ . \tag{257}$$

The remarkable thing about this way of giving a mass to the photon is that at no point have we 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 (256) we have not found any massless mode. Since the vacuum chosen by the scalar field breaks the $U(1)$ generator of U(1) we would have expected one massless particle from Goldstone's theorem. To understand the fate of the missing Goldstone boson we have to revisit the calculation leading to Eq. (256). 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 (256) 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 by 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 (256) 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's 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's 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 transformations 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:

$$\vec{A}_\mu(\vec{x})\longrightarrow\frac{1}{ig}g(\vec{x})^{-1}\vec{\nabla}g(\vec{x}), \qquad |\vec{x}|\to\infty\,. \tag{258}$$

The set transformations $g_0(\vec{x})\in\mathcal{G}_0$ that tend to the identity at infinity are the ones generated by Gauss's 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. Note 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's 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 $\mathrm{W}^{\pm}$ and Z^0 gauge bosons in the Standard Model. In condensed-matter physics the symmetry breaking described by the non-relativistic 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 non-zero mass of the photon (257) below the critical temperature explains the Meissner effect: the magnetic fields cannot penetrate inside superconductors beyond a distance of the order $1/m_\gamma$.

6 Anomalies

So far we have not worried 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 assumption, however, is not necessarily justified in the case of certain symmetries like scale invariance. To be more concrete, let us think of a theory containing a single field with canonical dimension Δ. If there are no dimensionful parameters in the Lagrangian, the classical theory is invariant under the conformal transformation

$$x^\mu \longrightarrow \lambda x^\mu\,, \qquad \phi(x) \longrightarrow \lambda^{-\Delta}\phi(\lambda^{-1}x)\,. \tag{259}$$

This is the case, for example, for a massless $\lambda\varphi^4$ theory in four dimensions

$$\mathcal{L} = \frac{1}{2}\partial_\mu\varphi\partial^\mu\varphi - \frac{\lambda}{4!}\varphi^4, \tag{260}$$

where the scalar field has canonical dimension $\Delta = 1$. The Lagrangian density transforms as

$$\mathcal{L} \longrightarrow \lambda^{-4}\mathcal{L} \tag{261}$$

and the classical action remains invariant.

This classical invariance of the theory is, however, not preserved in the process of quantization. The reason lies in the necessity of making sense of divergent expressions that arise when calculating quantum corrections, as we shall explain in Section 7 in detail. Here suffice it to say that in order to regularize the divergent expressions it is necessary to introduce a cut-off at a given energy scale. This breaking of the invariance of the theory under conformal transformations is not recovered after renormalization has been carried out, and as a result the quantum properties of a theory like Eq. (260) depend on the energy scale at which the physical processes take place. One of the consequences is that the canonical dimension of the field also gets a correction $\Delta = 1 + \gamma(\lambda)$.

This is an example of an *anomaly*, i.e., a symmetry of the classical theory that is not preserved upon quantization (for a review see Ref. [25]). 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. This is the case with the conformal anomaly that we have just discussed: It does not matter in which way we regularize our $\lambda\varphi^4$ theory, the result is a quantum theory that breaks conformal invariance.

6.1 Axial anomaly

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

$$\mathcal{L} = i\overline{\psi}\partial\!\!\!/\psi = iu_+^\dagger\sigma_+^\mu\partial_\mu u_+ + iu_-^\dagger\sigma_-^\mu\partial_\mu u_- \quad \text{with} \quad \psi = \begin{pmatrix} u_+ \\ u_- \end{pmatrix} \tag{262}$$

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:

$$\mathrm{U}(1)_V : u_\pm \longrightarrow e^{i\alpha}u_\pm\,. \tag{263}$$

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

$$\mathrm{U}(1)_A : u_\pm \longrightarrow e^{\pm i\alpha}u_\pm\,. \tag{264}$$

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

$$J_V^\mu = \overline{\psi}\gamma^\mu\psi = u_+^\dagger\sigma_+^\mu u_+ + u_-^\dagger\sigma_-^\mu u_- \qquad \Longrightarrow \qquad \partial_\mu J_V^\mu = 0 \tag{265}$$

and an axial vector current

$$J_A^\mu = \overline{\psi}\gamma^\mu\gamma_5\psi = u_+^\dagger\sigma_+^\mu u_+ - u_-^\dagger\sigma_-^\mu u_- \qquad \Longrightarrow \qquad \partial_\mu J_A^\mu = 0\,. \tag{266}$$

The theory described by the Lagrangian (262) can be coupled to the electromagnetic field. The resulting classical theory is still invariant under the vector and axial U(1) symmetries (263) and (264). Surprisingly, upon quantization it turns out that the conservation of the axial current (266) is spoiled by quantum effects:

$$\partial_\mu J_A^\mu \sim \hbar\,\vec{E}\cdot\vec{B}\,. \tag{267}$$

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 [26]. 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 set-up we consider a fermion coupled to the electromagnetic field. Note that since we are living in two dimensions, the field strength $F_{\mu\nu}$ has only 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

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu} \quad \text{with} \quad \eta = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\,. \tag{268}$$

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

$$\gamma^0 \equiv \sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \gamma^1 \equiv i\sigma^2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\,. \tag{269}$$

This is a chiral representation since the matrix γ_5 is diagonal[11],

$$\gamma_5 \equiv -\gamma^0\gamma^1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\,. \tag{270}$$

Writing the two-component spinor ψ as

$$\psi = \begin{pmatrix} u_+ \\ u_- \end{pmatrix} \tag{271}$$

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 a 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

$$(\partial_0 - \partial_1)u_+ = 0\,, \qquad (\partial_0 + \partial_1)u_- = 0\,. \tag{272}$$

The general solution to these equations can be immediately written as

$$u_+ = u_+(x^0 + x^1)\,, \qquad u_- = u_-(x^0 - x^1)\,. \tag{273}$$

Hence $u_\pm$ are two wave packets moving along the spatial dimension to the left (u_+) and to the right (u_-), respectively. Note 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).

[11] In any even number of dimensions γ_5 is defined to satisfy the conditions $\gamma_5^2 = \mathbf{1}$ and $\{\gamma_5, \gamma^\mu\} = 0$.

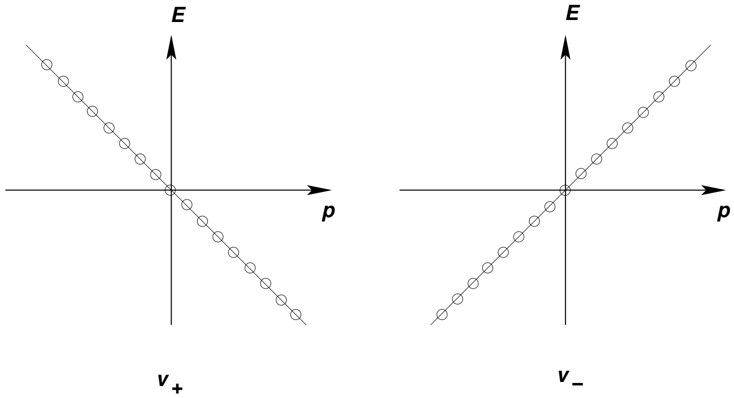


Fig. 10: Spectrum of the massless two-dimensional Dirac field

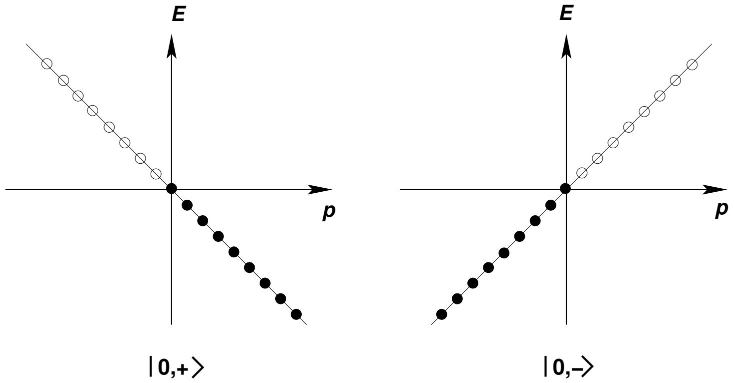


Fig. 11: Vacuum of the theory

If we want to interpret Eq. (272) as the wave equation for two-dimensional Weyl spinors we have the following wave functions for free particles with well defined momentum $p^\mu = (E,p)$:

$$u_\pm^{(E)}(x^0 \pm x^1) = \frac{1}{\sqrt{L}} e^{-iE(x^0 \pm x^1)} \quad \text{with} \quad p = \mp E\,. \tag{274}$$

As 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

$$p = \frac{2\pi n}{L}\,, \qquad n \in \mathbb{Z}\,. \tag{275}$$

The spectrum of the theory is represented in Fig. 10.

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. 11). Excitation of a particle in the Dirac sea produces a positive-energy fermion plus a hole that is interpreted as an antiparticle. This gives us a clue to how to quantize the theory. In the expansion of the operator $u_\pm$ in terms of the modes (274) we associate positive energy states with annihilation operators whereas the states with negative energy are associated with creation operators for the corresponding antiparticle:

$$u_\pm(x) = \sum_{E>0} \left[a_\pm(E) v_\pm^{(E)}(x) + b_\pm^\dagger(E) v_\pm^{(E)}(x)^* \right]\,. \tag{276}$$

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 $b^\dagger_\pm(E)$ 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. (276) satisfy the usual algebra

$$\{a_\lambda(E), a^\dagger_{\lambda'}(E')\} = \{b_\lambda(E), b^\dagger_{\lambda'}(E')\} = \delta_{E,E'}\delta_{\lambda\lambda'} \ , \tag{277}$$

where we have introduced the label $\lambda, \lambda' = \pm$. Also, $a_\lambda(E)$, $a^\dagger_\lambda(E)$ anticommute with $b_{\lambda'}(E')$, $b^\dagger_{\lambda'}(E')$.

The Lagrangian of the theory

$$\mathcal{L} = iu^\dagger_+(\partial_0+\partial_1)u_+ + iu^\dagger_-(\partial_0-\partial_1)u_- \tag{278}$$

is invariant under both U(1)$_V$, Eq. (263), and U(1)$_A$, Eq. (264). The associated Noether currents are in this case

$$J^\mu_V = \begin{pmatrix} u^\dagger_+u_+ + u^\dagger_-u_- \\ -u^\dagger_+u_+ + u^\dagger_-u_- \end{pmatrix} , \qquad J^\mu_A = \begin{pmatrix} u^\dagger_+u_+ - u^\dagger_-u_- \\ -u^\dagger_+u_+ - u^\dagger_-u_- \end{pmatrix} . \tag{279}$$

The associated conserved charges are given, for the vector current by

$$Q_V = \int_0^L dx^1 \left(u^\dagger_+u_+ + u^\dagger_-u_-\right) \tag{280}$$

and for the axial current by

$$Q_A = \int_0^L dx^1 \left(u^\dagger_+u_+ - u^\dagger_-u_-\right) . \tag{281}$$

Using the orthonormality relations for the modes $v^{(E)}_\pm(x)$

$$\int_0^L dx^1\, v^{(E)}_\pm(x)\, v^{(E')}_\pm(x) = \delta_{E,E'} \ , \tag{282}$$

we find for the conserved charges

$$\begin{aligned} Q_V &= \sum_{E>0}\Big[a^\dagger_+(E)a_+(E) - b^\dagger_+(E)b_+(E) + a^\dagger_-(E)a_-(E) - b^\dagger_-(E)b_-(E)\Big] , \\ Q_A &= \sum_{E>0}\Big[a^\dagger_+(E)a_+(E) - b^\dagger_+(E)b_+(E) - a^\dagger_-(E)a_-(E) + b^\dagger_-(E)b_-(E)\Big] . \end{aligned} \tag{283}$$

We see that Q_V counts the net number (particles minus antiparticles) of positive helicity states plus the net number of states with negative helicity. The axial charge, on the other hand, counts the net number of positive helicity states minus the number of negative helicity ones. In the case of the vector current we have subtracted a formally divergent vacuum contribution to the charge (the 'charge of the Dirac sea').

In the free theory there is of course no problem with the conservation of either Q_V or Q_A, since the occupation numbers do not change. What we want to study is the effect of coupling the theory to electric field $\mathcal{E}$. We work in the gauge $A_0 = 0$. Instead of solving the problem exactly we are going to simulate the electric field by adiabatically varying in a long time τ_0 the vector potential A_1 from zero value to

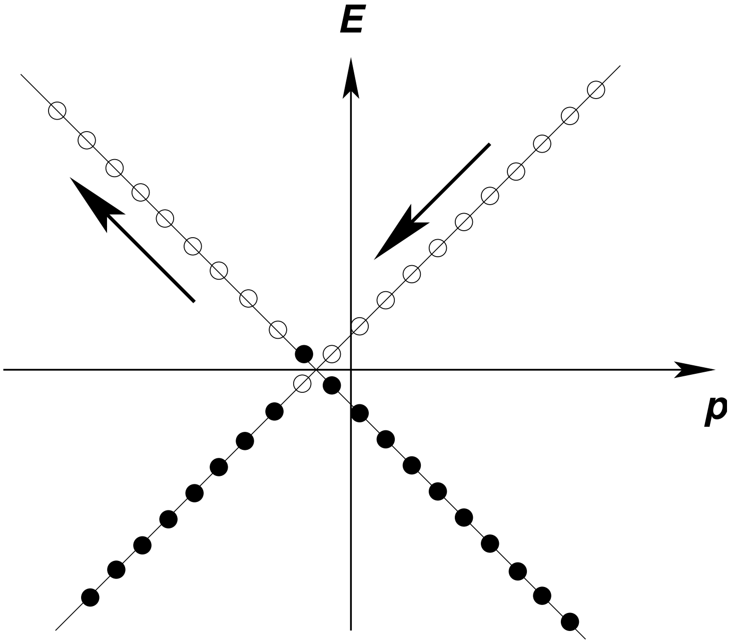


Fig. 12: Effect of the electric field

$-\mathcal{E}\tau_0$. 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

$$p \longrightarrow p - eA_1 \ , \tag{284}$$

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 Eq. (284) on the vacuum depicted in Fig. 11. What we find is that the two branches move as shown in Fig. 12, 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

$$\begin{aligned} Q_A(\tau_0) &= (N-0)+(0-N)=0 \ , \\ Q_V(\tau_0) &= (N-0)-(0-N)=2N \ . \end{aligned} \tag{285}$$

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

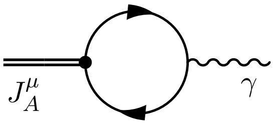
$$\partial_\mu J_A^\mu \sim e\hbar\mathcal{E} \ , \tag{286}$$

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, $\partial_\mu J_V^\mu = 0$.

We have just studied a two-dimensional example of the Adler–Bell–Jackiw axial anomaly [27].

The heuristic analysis presented here can be made more precise by computing the quantity

$$C^{\mu\nu} = \langle 0|T\left[J_A^\mu(x)J_V^\nu(0)\right]|0\rangle = \tag{287}$$



The anomaly is given then by $\partial_\mu C^{\mu\nu}$. A careful calculation yields the numerical prefactor missing in Eq. (286) leading to the result

$$\partial_\mu J_A^\mu = \frac{e\hbar}{2\pi}\varepsilon^{\nu\sigma}F_{\nu\sigma}\ , \tag{288}$$

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 space–times. In particular in four dimensions the axial anomaly is given by

$$\partial_\mu J_A^\mu = -\frac{e^2}{16\pi^2}\varepsilon^{\mu\nu\sigma\lambda}F_{\mu\nu}F_{\sigma\lambda}\ . \tag{289}$$

This result has very important consequences in the physics of strong interactions as we shall see in what follows.

6.2 Chiral symmetry in QCD

Our knowledge of the physics of strong interactions is based on the theory of quantum chromodynamics (QCD) [28]. This is a non-Abelian 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: colour $i = 1,\ldots,N_c$ and flavour $f = 1,\ldots,N_f$. 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 flavours 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 colours and N_f flavours. 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

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4}F_{\mu\nu}^a F^{a\,\mu\nu} + \sum_{f=1}^{N_f}\left[i\overline{Q}_L^f \not{D} Q_L^f + i\overline{Q}_R^f \not{D} Q_R^f\right]\ , \tag{290}$$

where the subscripts L and R indicate left- and right-handed spinors, respectively, $Q_{L,R}^f \equiv P_\pm Q^f$, and the field strength $F_{\mu\nu}^a$ and the covariant derivative D_μ are defined in Eqs. (165) and (168), respectively. Apart from the gauge symmetry, this Lagrangian is also invariant under a global U(N_f)$_L$ × U(N_f)$_R$ acting on the flavour indices and defined by

$$\text{U}(N_f)_L : \left\{\begin{array}{lll} Q_L^f & \to & \sum_{f'}(U_L)_{ff'}Q_L^{f'} \\ & & \\ Q_R^f & \to & Q_R^f \end{array}\right. \qquad \text{U}(N_f)_R : \left\{\begin{array}{lll} Q_L^f & \to & Q_L^f \\ & & \\ Q_R^r & \to & \sum_{f'}(U_R)_{ff'}Q_R^{f'} \end{array}\right. \tag{291}$$

with $U_L, U_R \in$ U(N_f). Actually, since U(N) = U(1) × SU(N) this global symmetry group can be written as SU(N_f)$_L$ × SU(N_f)$_R$ × U(1)$_L$ × U(1)$_R$. The Abelian subgroup U(1)$_L$ × U(1)$_R$ can now be decomposed into its vector $\mathrm{U}(1)_B$ and axial $\mathrm{U}(1)_A$ subgroups defined by the transformations

$$\mathrm{U}(1)_B : \left\{ \begin{array}{ccc} Q^f_L & \rightarrow & e^{i\alpha}Q^f_L \\ & & \\ Q^f_R & \rightarrow & e^{i\alpha}Q^f_R \end{array} \right. \qquad \mathrm{U}(1)_A : \left\{ \begin{array}{ccc} Q^f_L & \rightarrow & e^{i\alpha}Q^f_L \\ & & \\ Q^f_R & \rightarrow & e^{-i\alpha}Q^f_R \end{array} \right. \tag{292}$$

According to Noether's theorem, associated with these two Abelian symmetries we have two conserved currents:

$$J^\mu_V = \sum_{f=1}^{N_f} \overline{Q}^f \gamma^\mu Q^f \;, \qquad J^\mu_A = \sum_{f=1}^{N_f} \overline{Q}^f \gamma^\mu \gamma_5 \, Q^f \;. \tag{293}$$

The conserved charge associated with vector charge J^μ_V is actually the baryon number, defined as the number of quarks minus the number of antiquarks.

The non-Abelian 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 quark fields:

$$\mathrm{SU}(N_f)_V : \left\{ \begin{array}{ccc} Q^f_L & \rightarrow & \sum_{f'}(U_L)_{ff'}Q^{f'}_L \\ & & \\ Q^f_R & \rightarrow & \sum_{f'}(U_L)_{ff'}Q^{f'}_R \end{array} \right. \qquad \mathrm{SU}(N_f)_A : \left\{ \begin{array}{ccc} Q^f_L & \rightarrow & \sum_{f'}(U_L)_{ff'}Q^{f'}_L \\ & & \\ Q^f_R & \rightarrow & \sum_{f'}(U^{-1}_R)_{ff'}Q^{f'}_R \end{array} \right. \tag{294}$$

Again, the application of Noether's theorem shows the existence of the following non-Abelian conserved charges:

$$J^{I\,\mu}_V \equiv \sum_{f,f'=1}^{N_f} \overline{Q}^f \gamma^\mu (T^I)_{ff'} Q^{f'} \;, \qquad J^{I\,\mu}_A \equiv \sum_{f,f'=1}^{N_f} \overline{Q}^f \gamma^\mu \gamma_5 (T^I)_{ff'} Q^{f'} \;. \tag{295}$$

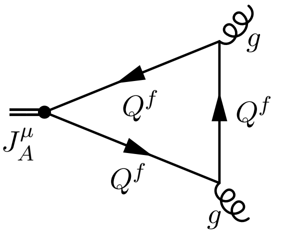
To summarize, we have shown that the initial chiral symmetry of the QCD Lagrangian (290) can be decomposed into its chiral and vector subgroups according to

$$\mathrm{U}(N_f)_L \times \mathrm{U}(N_f)_R = \mathrm{SU}(N_f)_V \times \mathrm{SU}(N_f)_A \times \mathrm{U}(1)_B \times \mathrm{U}(1)_A \;. \tag{296}$$

The question to address now is which part of the classical global symmetry is preserved by the quantum theory.

As argued in Section 6.1, the conservation of the axial currents J^μ_A and $J^{a\,\mu}_A$ can in principle be spoiled due to the presence of an anomaly. In the case of the Abelian axial current J^μ_A the relevant quantity is the correlation function

$$C^{\mu\nu\sigma} \equiv \langle 0|T\left[J^\mu_A(x) j^{a\,\nu}_{\mathrm{gauge}}(x') j^{b\,\sigma}_{\mathrm{gauge}}(0)\right]|0\rangle = \sum_{f=1}^{N_f} \Big[\qquad \Big]_{\mathrm{symmetric}} \tag{297}$$



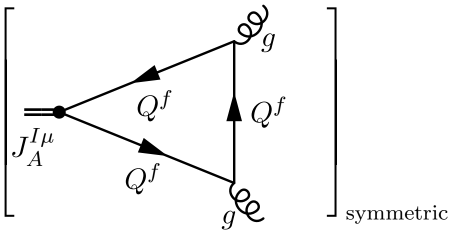
Here $j^{a\,\mu}_{\mathrm{gauge}}$ is the non-Abelian conserved current coupling to the gluon field

$$j^{a\,\mu}_{\mathrm{gauge}} \equiv \sum_{f=1}^{N_f} \overline{Q}^f \gamma^\mu \tau^a Q^f \;, \tag{298}$$

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 now be read 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[12]

$$\partial_\mu J_A^\mu = -\frac{g^2 N_f}{32\pi^2}\varepsilon^{\mu\nu\sigma\lambda}F^a_{\mu\nu}F^{a\,\mu\nu}\ . \tag{299}$$

In the case of the non-Abelian 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 non-Abelian axial current $J_A^{a\,\mu}$ is not anomalous. This can easily be seen by noting 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 theoretical factor,

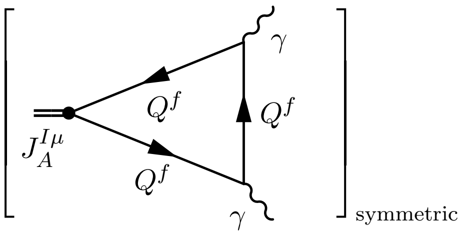


$$\sim \text{tr}\, T^I\ \text{tr}\,\{\tau^a,\tau^b\} = 0\ , \tag{300}$$

which vanishes because the generators of SU(N_f) are traceless.

From here we would conclude that the non-Abelian 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

$$\langle 0|T\left[J_A^{I\,\mu}(x)j^\nu_{\rm em}(x')j^\sigma_{\rm em}(0)\right]|0\rangle = \sum_{f=1}^{N_f} \tag{301}$$



where $j^\mu_{\rm em}$ is the electromagnetic current

$$j^\mu_{\rm em} = \sum_{f=1}^{N_f} q_f\,\overline{Q}^f\gamma^\mu Q^f\ , \tag{302}$$

with q_f the electric charge of the f-th quark flavour. A calculation of the diagram in Eq. (301) shows the existence of an Adler–Bell–Jackiw anomaly given by

$$\partial_\mu J_A^{I\,\mu} = -\frac{N_c}{16\pi^2}\left[\sum_{f=1}^{N_f}(T^I)_{ff}\,q_f^2\right]\varepsilon^{\mu\nu\sigma\lambda}F_{\mu\nu}F_{\sigma\lambda}\ , \tag{303}$$

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.

[12] The normalization of the generators T^I of the global SU(N_f) is given by tr $(T^IT^J) = \frac{1}{2}\delta^{IJ}$.

Before proceeding let us summarize the results found so far. Because of the presence of anomalies, the axial part of the global chiral symmetry, SU(N_f)$_A$ and U(1)$_A$, are not realized quantum mechanically in general. We found that U(1)$_A$ is always affected by an anomaly. However, because the right-hand side of the anomaly equation (299) is a total derivative, the anomalous character of J_A^μ does not explain the absence of U(1)$_A$ multiplets in the hadron spectrum, since a new current can be constructed which is conserved. In addition, the non-existence of candidates for a Goldstone boson associated with the right quantum numbers indicates that U(1)$_A$ is not spontaneously broken either, so it has to be explicitly broken somehow. This is the so-called U(1)-problem which was solved by 't Hooft [29], 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.

Because of the dynamics of the SU(N_c) gauge theory the axial non-Abelian symmetry is spontaneously broken due to the presence at low energies of a vacuum expectation value for the fermion bilinear $\overline{Q}^f Q^f$:

$$\langle 0|\overline{Q}^f Q^f|0\rangle \neq 0 \qquad\qquad \text{(No summation in } f\text{!)} . \tag{304}$$

This non-vanishing vacuum expectation value for the quark bilinear actually breaks chiral invariance spontaneously to the vector subgroup SU(N_f)$_V$, so the only subgroup of the original global symmetry that is realized by the full theory at low energy is

$$\text{U}(N_f)_L \times \text{U}(N_f)_R \longrightarrow \text{SU}(N_f)_V \times \text{U}(1)_B . \tag{305}$$

Associated with this breaking, a Goldstone boson should appear with the quantum numbers of the broken non-Abelian current. For example, in the case of QCD the Goldstone bosons associated with the spontaneous 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 non-vanishing 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 with 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 flavours, QCD at low energies can be well described by including only these two flavours and ignoring heavier quarks. In this approximation, from our previous discussion we know that the low-energy global symmetry of the theory is SU(2)$_V$×U(1)$_B$, where now the vector group SU(2)$_V$ is the well-known isospin symmetry. The axial U(1)$_A$ current is anomalous due to Eq. (299) with $N_f = 2$. In the case of the non-Abelian axial symmetry SU(2)$_A$, taking into account that $q_u = \frac{2}{3}e$ and $q_d = -\frac{1}{3}e$ and that the three generators of SU(2) can be written in terms of the Pauli matrices as $T^K = \frac{1}{2}\sigma^K$, we find

$$\sum_{f=u,d}(T^1)_{ff}\, q_f^2 = \sum_{f=u,d}(T^1)_{ff}\, q_f^2 = 0 \,, \qquad \sum_{f=u,d}(T^3)_{ff}\, q_f^2 = \frac{e^2}{6} \,. \tag{306}$$

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 quarks by

$$|\pi^0\rangle = \frac{1}{\sqrt{2}}\big(|\bar{u}\rangle|u\rangle - |\bar{d}\rangle|d\rangle\big) \ . \tag{307}$$

The isospin quantum numbers of $|\pi^0\rangle$ are those of the generator T^3. In fact the analogy goes further since $\partial_\mu J_A^{3\,\mu}$ is the operator creating a pion π^0 out of the vacuum,

$$|\pi^0\rangle \sim \partial_\mu J_A^{3\,\mu}|0\rangle \ . \tag{308}$$

This leads to the physical interpretation of the triangle diagram (301) with $J_A^{3\,\mu}$ as the one-loop contribution to the decay of a neutral pion into two photons,

$$\pi^0 \longrightarrow 2\gamma \,. \tag{309}$$

This is an interesting piece of physics. In 1967 Sutherland and Veltman [30] 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 of this paradox, as pointed out in Ref. [27], 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 non-vanishing value for the $\pi^0 \to 2\gamma$ amplitude[13].

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 that 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 non-decoupling 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

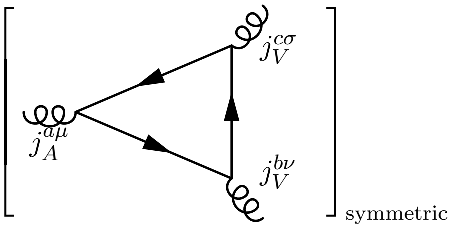
$$\mathcal{L} = -\frac{1}{4}F^{a\,\mu\nu}F^a_{\mu\nu} + i\sum_{i=1}^{N_+}\overline{\psi}{}^i_+ \not{D}^{(+)}\psi^i_+ + i\sum_{j=1}^{N_-}\overline{\psi}{}^j_- \not{D}^{(-)}\psi^j_- \,, \tag{310}$$

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

$$D^{(\pm)}_\mu \psi^i_\pm = \partial_\mu \psi^i_\pm + igA^K_\mu \tau^K_{i,\pm}\psi^i_\pm \,. \tag{311}$$

As for global symmetries, anomalies in the gauge symmetry appear in the triangle diagram with one axial and two vector gauge current vertices,

$$\langle 0|T\left[j_A^{a\,\mu}(x)j_V^{b\,\nu}(x')j_V^{c\,\sigma}(0)\right]|0\rangle = \tag{312}$$



[13] An early computation of the triangle diagram for the electromagnetic decay of the pion was made by Steinberger [31].

where gauge vector and axial currents $j_V^{a\,\mu}$, $j_A^{a\,\mu}$ are given by

$$\begin{aligned} j_V^{a\mu} &= \sum_{i=1}^{N_+} \overline{\psi}_+^i \tau_+^a \gamma^\mu \psi_+^i + \sum_{j=1}^{N_-} \overline{\psi}_-^j \tau_-^a \gamma^\mu \psi_-^j \ , \\ j_A^{a\mu} &= \sum_{i=1}^{N_+} \overline{\psi}_+^i \tau_+^a \gamma^\mu \psi_+^i - \sum_{i=1}^{N_-} \overline{\psi}_-^j \tau_-^a \gamma^\mu \psi_-^j \ . \end{aligned} \tag{313}$$

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. (312) for every fermion species running in the loop this factor is equal to

$$\mathrm{tr}\,\Big[\tau_{i,\pm}^a\{\tau_{i,\pm}^b, \tau_{i,\pm}^c\}\Big] \ , \tag{314}$$

where the sign ± corresponds to the generators of the representation of the gauge group for the left- and right-handed fermions, respectively. Hence the anomaly cancellation condition reads

$$\sum_{i=1}^{N_+} \mathrm{tr}\,\Big[\tau_{i,+}^a\{\tau_{i,+}^b, \tau_{i,+}^c\}\Big] - \sum_{j=1}^{N_-} \mathrm{tr}\,\Big[\tau_{j,-}^a\{\tau_{j,-}^b, \tau_{j,-}^c\}\Big] = 0 \ . \tag{315}$$

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

$$Q = T_3 + Y \ , \tag{316}$$

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:

$$\begin{array}{llll} \text{quarks:} & \begin{pmatrix} u^\alpha \\ d^\alpha \end{pmatrix}_{L,\frac{1}{6}} & u^\alpha_{R,\frac{2}{3}} & d^\alpha_{R,\frac{2}{3}} \\ \text{leptons:} & \begin{pmatrix} \nu_e \\ e \end{pmatrix}_{L,-\frac{1}{2}} & e_{R,-1} & \end{array} \tag{317}$$

where $\alpha = 1, 2, 3$ labels the colour 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:

$$\begin{array}{llll} \text{left-handed fermions:} & (3,2)^L_{\frac{1}{6}} & (1,2)^L_{-\frac{1}{2}} & \\ \text{right-handed fermions:} & (3,1)^R_{\frac{2}{3}} & (3,1)^R_{-\frac{1}{3}} & (1,1)^R_{-1} \ . \end{array} \tag{318}$$

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:

SU(3)^3

SU(3)^2 SU(2)

SU(3)^2 U(1)

SU(3) SU(2)^2

SU(3) SU(2) U(1)

SU(3) U(1)^2

SU(2)^3

SU(2)^2 U(1)

SU(2) U(1)^2

U(1)^3

It is easy to check that some of them do not give rise to anomalies. For example the anomaly for the SU(3)^3 case cancels because left- and right-handed quarks transform in the same representation. In the case of SU(2)^3 the cancellation happens term by term because of the Pauli matrices identity $\sigma^a\sigma^b = \delta^{ab} + i\varepsilon^{abc}\sigma^c$ that leads to

$$\operatorname{tr}\left[\sigma^a\{\sigma^b,\sigma^c\}\right] = 2\,(\operatorname{tr}\sigma^a)\,\delta^{bc} = 0\,. \tag{319}$$

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 non-trivial identity

$$\begin{aligned}\sum_{\text{left}} Y_+^3 - \sum_{\text{right}} Y_-^3 &= 3\times 2\times\left(\frac{1}{6}\right)^3 + 2\times\left(-\frac{1}{2}\right)^3 - 3\times\left(\frac{2}{3}\right)^3 - 3\times\left(-\frac{1}{3}\right)^3 - (-1)^3 \\ &= \left(-\frac{3}{4}\right) + \left(\frac{3}{4}\right) = 0\,. \end{aligned}\tag{320}$$

It is remarkable that the anomaly exactly cancels between leptons and quarks. Note 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 [32] guarantees that the axial anomaly only receives contributions from one-loop diagrams. Therefore once anomalies are cancelled (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, applies only in perturbation theory. It is nonetheless possible that non-perturbative effects can result in the quantum violation of a gauge symmetry. This is precisely the case pointed out by Witten [33] with respect to the SU(2) gauge symmetry of the Standard Model. In this case the problem lies in the non-trivial 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 does gauge invariance allow for a non-trivial theory. It is again remarkable that the family structure of the Standard Model makes this anomaly cancel:

$$3\times\begin{pmatrix}u\\ d\end{pmatrix}_L + 1\times\begin{pmatrix}\nu_e\\ e\end{pmatrix}_L = 4\ \text{SU(2)-doublets}\,, \tag{321}$$

where the factor of 3 comes from the number of colours.

7 Renormalization

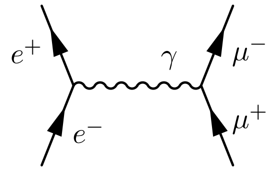
7.1 Removing infinities

From its very early stages, quantum field theory was faced with infinities. They emerged in the calculation of most physical quantities, such as the correction to the charge of the electron due to the interactions with the radiation field. The way these divergences were handled in the 1940s, starting with Kramers, was physically very much in the spirit of the quantum theory emphasis in observable quantities: Since the observed magnitude of physical quantities (such as the charge of the electron) is finite, this number should arise from the addition of a 'bare' (unobservable) value and the quantum corrections. The fact that both of these quantities were divergent was not a problem physically, since only its finite sum was an observable quantity. To make things mathematically sound, the handling of infinities requires the introduction of some regularization procedure which cuts the divergent integrals off at some momentum scale Λ. Morally speaking, the physical value of an observable $\mathcal{O}_{\rm physical}$ is given by

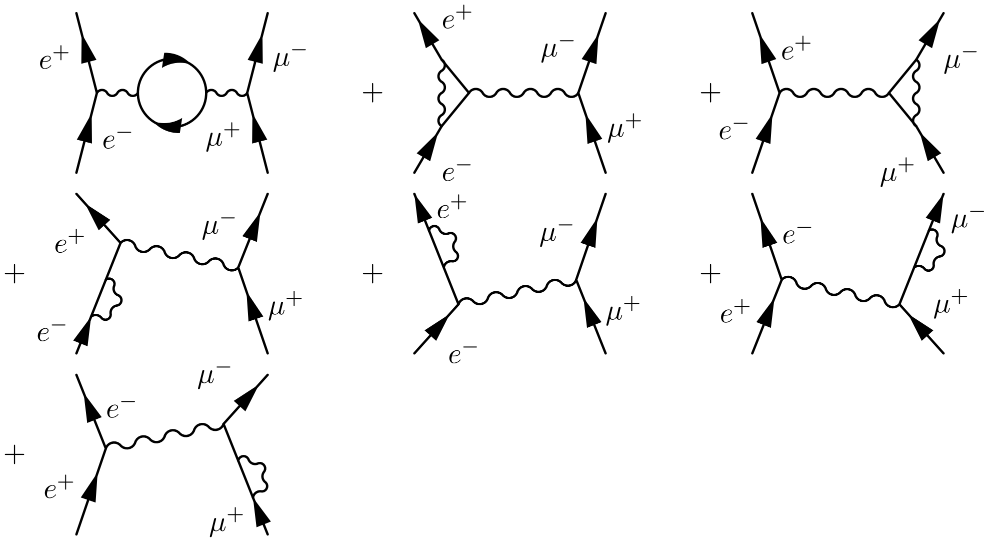
$$\mathcal{O}_{\rm physical} = \lim_{\Lambda\to\infty}\left[\mathcal{O}(\Lambda)_{\rm bare} + \Delta\mathcal{O}(\Lambda)_{\hbar}\right] , \qquad (322)$$

where $\Delta\mathcal{O}(\Lambda)_{\hbar}$ represents the regularized quantum corrections.

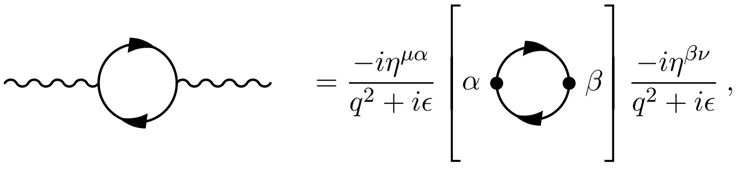
To make this qualitative discussion more precise we compute the corrections to the electric charge in quantum electrodynamics. We consider the process of annihilation of an electron–positron pair to create a muon–antimuon pair $e^-e^+ \to \mu^+\mu^-$. To lowest order in the electric charge e the only diagram contributing is



However, the corrections at order e^4 to this result requires the calculation of seven more diagrams:

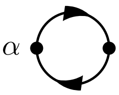


In order to compute the renormalization of the charge we consider the first diagram which takes into account the first correction to the propagator of the virtual photon interchanged between the pairs due to vacuum polarization. We begin by evaluating



$$= \frac{-i\eta^{\mu\alpha}}{q^2+i\epsilon}\left[\alpha \qquad \beta\right]\frac{-i\eta^{\beta\nu}}{q^2+i\epsilon} , \qquad (323)$$

where the diagram between brackets is given by

α  β

$$\equiv \Pi^{\alpha\beta}(q) = i^2(-ie)^2(-1)\int \frac{d^4k}{(2\pi)^4}\frac{\mathrm{Tr}\,(\not{k}+m_e)\gamma^\alpha(\not{k}+\not{q}+m_e)\gamma^\beta}{[k^2-m_e^2+i\epsilon]\,[(k+q)^2-m_e^2+i\epsilon]}\,. \tag{324}$$

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 of a photon and the creation of an electron–positron pair or the creation of a photon and annihilation of an electron–positron pair. This can be implemented by the interaction Hamiltonian

$$H_{\rm int} = e\int d^3x\,\overline{\psi}\gamma^\mu\psi A_\mu\,. \tag{325}$$

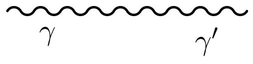
All fields inside the integral can be expressed in terms of the corresponding creation–annihilation operators for photons, electrons and positrons. In quantum mechanics, the change in the wave function at first order in the perturbation $H_{\rm int}$ is given by

$$|\gamma,\mathrm{in}\rangle = |\gamma,\mathrm{in}\rangle_0 + \sum_n \frac{\langle n|H_{\rm int}|\gamma,\mathrm{in}\rangle_0}{E_{\rm in}-E_n}|n\rangle \tag{326}$$

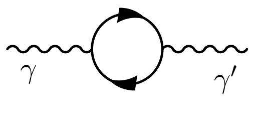
and similarly for $|\gamma,\mathrm{out}\rangle$, where we have denoted symbolically by $|n\rangle$ all the possible states of the electron–positron pair. Since these states are orthogonal to $|\gamma,\mathrm{in}\rangle_0$, $|\gamma,\mathrm{out}\rangle_0$, we find to order e^2

$$\langle\gamma,\mathrm{in}|\gamma',\mathrm{out}\rangle = {}_0\langle\gamma,\mathrm{in}|\gamma',\mathrm{out}\rangle_0 + \sum_n \frac{{}_0\langle\gamma,\mathrm{in}|H_{\rm int}|n\rangle\,\langle n|H_{\rm int}|\gamma',\mathrm{out}\rangle_0}{(E_{\rm in}-E_n)(E_{\rm out}-E_n)} + \mathcal{O}(e^4)\,. \tag{327}$$

Hence, we see that the diagram of Eq. (323) really corresponds to the order-e^2 correction to the photon propagator $\langle\gamma,\mathrm{in}|\gamma',\mathrm{out}\rangle$



$$\longrightarrow \quad {}_0\langle\gamma,\mathrm{in}|\gamma',\mathrm{out}\rangle_0$$



$$\longrightarrow \quad \sum_n \frac{\langle\gamma,\mathrm{in}|H_{\rm int}|n\rangle\,\langle n|H_{\rm int}|\gamma',\mathrm{out}\rangle}{(E_{\rm in}-E_n)(E_{\rm out}-E_n)}\,. \tag{328}$$

Once we have 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. (323) for non-zero 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. Doing some γ-matrices gymnastics it is not complicated to show that the polarization tensor $\Pi_{\mu\nu}(q)$ defined in Eq. (324) can be written as

$$\Pi_{\mu\nu}(q) = \left(q^2\eta_{\mu\nu} - q_\mu q_\nu\right)\Pi(q^2) \tag{329}$$

with

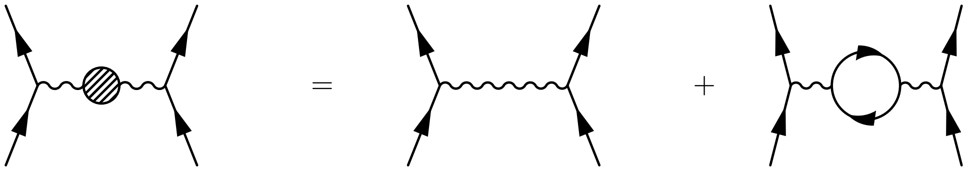
$$\Pi(q^2) = \frac{4e^2}{3q^2}\int \frac{d^4k}{(2\pi)^4}\frac{k^2 + k\cdot q}{[k^2+i\epsilon]\,[(k+q)^2+i\epsilon]}\ . \tag{330}$$

Although by naive power counting we could conclude that the previous integral is quadratically divergent, it can be seen that the quadratic divergence actually cancels leaving behind only a logarithmic one. In order to handle this divergent integral we have to work 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

$$\Pi(q^2) \simeq \frac{e^2}{12\pi^2}\log\left(\frac{q^2}{\Lambda^2}\right) + \text{finite terms}\ . \tag{331}$$

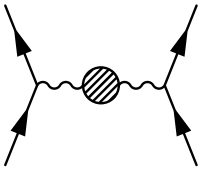
If we send the cut-off to infinity $\Lambda \to \infty$, the divergence blows up and something has to be done about it.

If we want to make sense out of this, we have to go back to the physical question that led us to compute Eq. (323). 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



$$\begin{aligned} &= \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\frac{e^2}{4\pi q^2}\left(\overline{v}_\mu\gamma^\beta u_\mu\right) + \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\frac{e^2}{4\pi q^2}\Pi(q^2)\left(\overline{v}_\mu\gamma^\beta u_\mu\right) \\ &= \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\left\{\frac{e^2}{4\pi q^2}\left[1+\frac{e^2}{12\pi^2}\log\left(\frac{q^2}{\Lambda^2}\right)\right]\right\}\left(\overline{v}_\mu\gamma^\beta u_\mu\right)\ . \end{aligned} \tag{332}$$

Now let us imagine that we are performing a $e^-\,e^+ \to \mu^-\mu^+$ with a centre-of-mass energy μ. From the previous result we can identify the effective charge of the particles at this energy scale $e(\mu)$ as



$$= \quad \eta_{\alpha\beta}\left(\overline{v}_e\gamma^\alpha u_e\right)\left[\frac{e(\mu)^2}{4\pi q^2}\right]\left(\overline{v}_\mu\gamma^\beta u_\mu\right)\ . \tag{333}$$

This charge, $e(\mu)$, is the quantity that is physically measurable in our experiment. Now we can make sense of the formally divergent result (332) 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 Eq. (332) with the physical results (333) we must assume that the dependence of the bare (unobservable) charge $e(\Lambda)_{\text{bare}}$ on the cut-off Λ is determined by the identity

$$e(\mu)^2 = e(\Lambda)^2_{\text{bare}}\left[1+\frac{e(\Lambda)^2_{\text{bare}}}{12\pi^2}\log\left(\frac{\mu^2}{\Lambda^2}\right)\right]\ . \tag{334}$$

If we still insist on removing the cut-off $\Lambda \to \infty$, we have to send the bare charge to zero $e(\Lambda)_{\text{bare}} \to 0$ 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 cut-off, since the only measurable quantity is the effective charge that remains finite. Therefore all observable quantities should be expressed in perturbation theory as a power series in the physical coupling $e(\mu)^2$ and not in the unphysical bare coupling $e(\Lambda)_{\text{bare}}$.

7.2 The beta function and asymptotic freedom

We can look at the previous discussion, and in particular Eq. (334), 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 (334) we can actually eliminate the cut-off 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 $e(\Lambda)^2_{\text{bare}}$ in terms of $e(\mu)^2$ as

$$e(\Lambda)^2 = e(\mu)^2\left[1+\frac{e(\mu)^2}{12\pi^2}\log\left(\frac{\mu^2}{\Lambda^2}\right)\right]+\mathcal{O}[e(\mu)^6]\,. \tag{335}$$

This expression allows us to eliminate all dependence on the cutoff in the expression of the effective charge at a scale μ by replacing $e(\Lambda)_{\text{bare}}$ in Eq. (334) with the one computed using Eq. (335) at a given reference energy scale μ_0:

$$e(\mu)^2 = e(\mu_0)^2\left[1+\frac{e(\mu_0)^2}{12\pi^2}\log\left(\frac{\mu^2}{\mu_0^2}\right)\right]. \tag{336}$$

From this expression 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, where the value of the electron charge is given by the well-known value

$$e(1\,\text{eV})^2 \simeq \frac{1}{137}\,. \tag{337}$$

Knowing this we can compute $e(\mu)^2$ at any other energy scale, for example at the mass of the Z^0 boson $\mu = M_{\text{Z}} \equiv 92$ GeV

$$e(M_{\text{Z}})^2 = e(m_e)^2\left[1+\frac{e(m_e)^2}{12\pi^2}\log\left(\frac{M_{\text{Z}}^2}{m_e^2}\right)\right]\simeq\frac{1}{128}\,. \tag{338}$$

Therefore we find that the electromagnetic coupling grows with energy. This can be explained heuristically by remembering that the effect of the polarization of the vacuum shown in the diagram of Eq. (323) 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 decrease 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

$$\beta(g) = \mu\frac{\mathrm{d}g}{\mathrm{d}\mu}\,. \tag{339}$$

In the case of QED the beta function can be computed from Eq. (336) with the result

$$\beta(e)_{\text{QED}} = \frac{e^3}{12\pi^2}\,. \tag{340}$$

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 behaviour of the coupling as we change the scale. Equation (340) 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. In particular this means that the coupling constant of the theory will tend to zero when the energy scale tends to zero

$$\lim_{\mu\rightarrow 0} e(\mu)^2 = 0\,, \tag{341}$$

so the perturbative approximation gives better and better results as we go to lower energies. 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 shall 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 behaviour 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 seemed 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 [34] and David Politzer [35] showed that non-Abelian gauge theories can actually display the required behaviour. For the QCD Lagrangian in Eq. (290) the beta function is given by[14]

$$\beta(g) = -\frac{g^3}{16\pi^2}\left[\frac{11}{3}N_c - \frac{2}{3}N_f\right] . \tag{342}$$

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 explains the apparent freedom of quarks inside the hadrons: when the quarks are very close together their effective colour charge tends to zero. This phenomenon is called *asymptotic freedom*.

Asymptotic free theories display a behaviour that is opposite to that found above in QED. At high energies their coupling constant approaches zero whereas at low energies they become strongly coupled (infrared slavery). These features are at the heart of the success of QCD as a theory of strong interactions, since this is exactly the type of behaviour found in quarks: They are quasi-free particles inside the hadrons but the interaction 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 colour charges and generates the spectrum of hadrons, as well as the breaking of the chiral symmetry (304).

7.3 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 on occasions by comparing it to ‘sweeping the infinities under the rug’. However, thanks to a large extent to Ken Wilson [38] 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.

[14] The expression of the beta function of QCD was also known to ’t Hooft [36]. There are even earlier computations in the Russian literature [37].

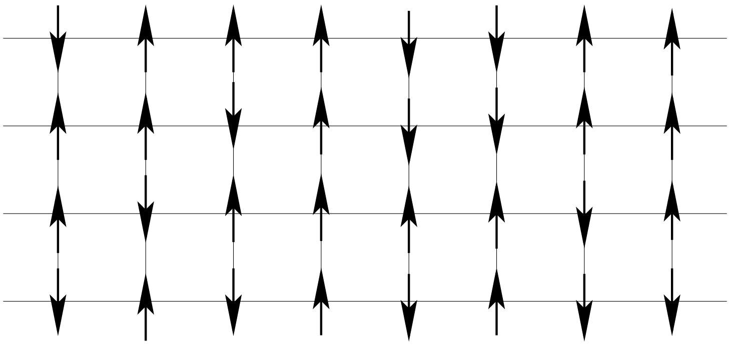


Fig. 13: Systems of spin in a two-dimensional square lattice

Statistical mechanics

Wilson's ideas are both simple and profound and consist in thinking about quantum field theory as the analogue 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 such as the one depicted in Fig. 13. 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

$$H = -J\sum_{\langle i,j\rangle} s_i\, s_j\ , \tag{343}$$

where $\langle i,j\rangle$ indicates that the sum extends over nearest neighbours and J is the coupling constant between neighbouring 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

$$\mathcal{Z} = \sum_{\{s_i\}} e^{-\beta H}\ , \tag{344}$$

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

$$\langle s_i s_j\rangle \sim e^{-\frac{|x_{ij}|}{\xi}}\ , \tag{345}$$

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 neighbours.

Suppose now that we are interested in a macroscopic description of this spin system. We can capture the relevant physics by somehow integrating out the physics at short scales. A way in which this can be done was proposed by Leo Kadanoff [39] and consists in dividing our spin system in spin-blocks like the ones shown in Fig. 14. 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:

$$\{s_i : i \in B_a\} \qquad \longrightarrow \qquad {s_a}^{(1)}\ . \tag{346}$$

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:

$${s_a}^{(1)} = \frac{1}{2}\text{sgn}\left(\sum_{i\in B_a} s_i\right)\ , \tag{347}$$

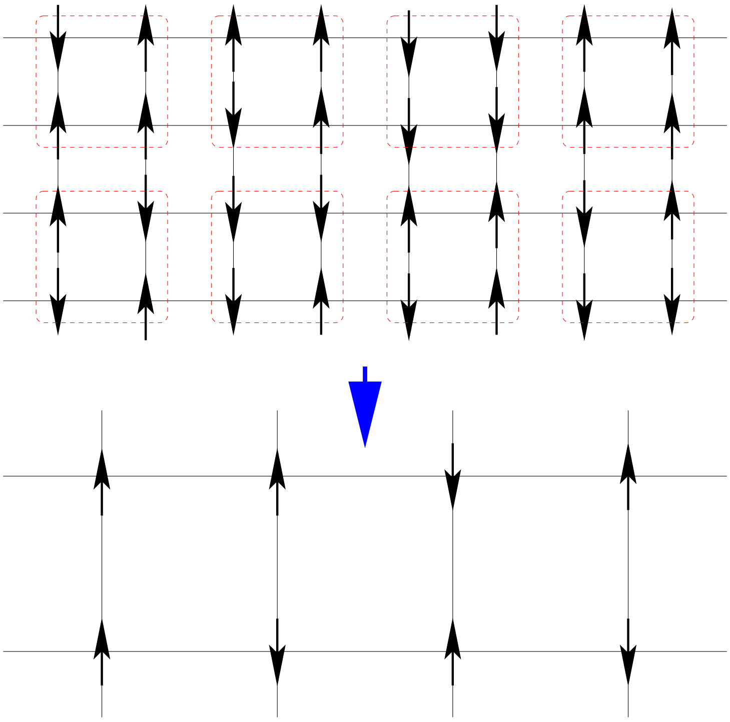


Fig. 14: Decimation of the spin lattice. Each block in the upper lattice is replaced by an effective spin computed according to the rule of Eq. (347). Note that the size of the lattice spacing is doubled in the process.

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 (344) 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

$$\mathcal{Z} = \sum_{\{\vec{s}\}} e^{-\beta H[s_i]} = \sum_{\{\vec{s}^{(1)}\}} \sum_{\{\vec{s}\in B_a\}} \delta\left[{s_a}^{(1)} - \text{sign}\left(\sum_{i\in B_a} s_i\right)\right] e^{-\beta H[s_i]}\,. \tag{348}$$

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: $H^{(1)}[{s_a}^{(1)}]$

$$\sum_{\{s\in B_a\}} \delta\left[{s_a}^{(1)} - \text{sign}\left(\sum_{i\in B_a} s_i\right)\right] e^{-\beta H[s_i]} = e^{-\beta H^{(1)}[{s_a}^{(1)}]}\,. \tag{349}$$

The new Hamiltonian is of course more complicated:

$$H^{(1)} = -J^{(1)} \sum_{\langle i,j\rangle} s_i^{(1)} s_j^{(1)} + \ldots\ , \tag{350}$$

where the dots stand for other interaction terms between the effective block spins. These 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 neighbours in the new lattice. The important point is that we have managed to rewrite the partition function solely in terms of these new (renormalized) spin variables

$s^{(1)}$ interacting through a new Hamiltonian $H^{(1)}$:

$$\mathcal{Z} = \sum_{\{s^{(1)}\}} e^{-\beta H^{(1)}[s_a^{(1)}]} \,. \tag{351}$$

Let us now think about the space of all possible Hamiltonians for our statistical system including all possible types of coupling between the individual spins compatible with the symmetries of the system. If we denote by $\mathcal{R}$ the decimation operation, our previous analysis shows that $\mathcal{R}$ defines a map in this space of Hamiltonians,

$$\mathcal{R} : H \to H^{(1)} \,. \tag{352}$$

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 $\mathcal{R}$,

$$H \xrightarrow{\mathcal{R}} H^{(1)} \xrightarrow{\mathcal{R}} H^{(2)} \xrightarrow{\mathcal{R}} \ldots \xrightarrow{\mathcal{R}} H_\star \,. \tag{353}$$

The fixed-point Hamiltonian $H_\star$ is *scale invariant* because it does not change as $\mathcal{R}$ is performed. Note that because of this invariance the correlation length of the system at the fixed point does 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 shall focus on the case of non-trivial 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

$$H[s_i] = \sum_{a=1}^{\infty} \lambda_a \mathcal{O}_a[s_i] \,, \tag{354}$$

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,

$$\mathcal{R} : \lambda_a \longrightarrow \lambda_a^{(1)} \,. \tag{355}$$

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 non-zero 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 results 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

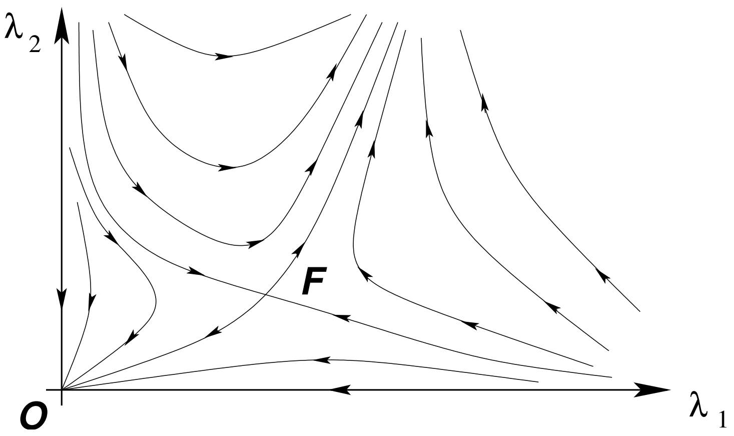


Fig. 15: Example of a renormalization group flow

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 replacing $H_\star$ by the perturbed Hamiltonian

$$H = H_\star + \delta\lambda\,\mathcal{O}\ , \tag{356}$$

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 decision to make 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 from 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. 15 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 fixed 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 a certain type of perturbation (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

We shall now see how these ideas of the renormalization group apply to field theory. Let us begin with a quantum field theory defined by the Lagrangian

$$\mathcal{L}[\phi_a] = \mathcal{L}_0[\phi_a] + \sum_i g_i \mathcal{O}_i[\phi_a] \ , \tag{357}$$

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 cut-off in momenta Λ. In principle we include all operators $\mathcal{O}_i$ compatible with the symmetries of the theory.

In Section 7.2 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 now understand this behaviour 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 cut-off Λ and the scale of interest μ. This is analogous to what we did in the Ising model by replacing the original spins with 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

$$e^{iS[\phi'_a,\mu]} = \int_{\mu<p<\Lambda} \prod_a \mathcal{D}\phi_a \, e^{iS[\phi_a,\Lambda]} \ . \tag{358}$$

Here $S[\phi_a, \Lambda]$ is the action at the cut-off scale,

$$S[\phi_a, \Lambda] = \int d^4x \left\{ \mathcal{L}_0[\phi_a] + \sum_i g_i(\Lambda) \mathcal{O}_i[\phi_a] \right\} \ , \tag{359}$$

and the functional integral in Eq. (358) 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 cut-off 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

$$S[\phi'_a, \mu] = \int d^4x \left\{ \mathcal{L}_0[\phi'_a] + \sum_i g_i(\mu) \mathcal{O}_i[\phi'_a] \right\} \ . \tag{360}$$

This Wilsonian interpretation of renormalization sheds light on what in Section 7.1 might have looked like just a clever 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 we can also find quantum field theories that are fixed points of the renormalization group flow, i.e., whose coupling constants do not change with the scale. 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 what happens when a scale-invariant theory is perturbed with some operator. In general the perturbed theory is not scale invariant any more, but we may wonder whether the perturbed theory flows at low energies towards or away from 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 the following:

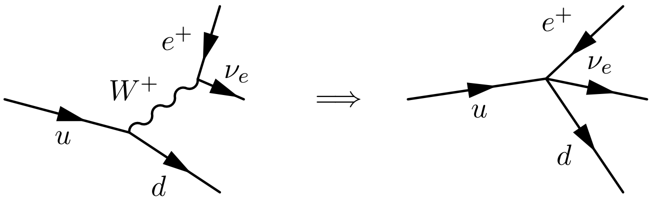
- $d[\mathcal{O}] > 4$: irrelevant perturbation. The running of the coupling constants takes the theory back to the fixed point.
- $d[\mathcal{O}] < 4$: relevant perturbation. At low energies the theory flows away from the scale-invariant theory.
- $d[\mathcal{O}] = 4$: marginal deformation. The direction 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:

$$\mathcal{L} = i\overline{\psi}\partial\!\!\!/\psi - \frac{1}{M^2}(\overline{\psi}\psi)^2 \; . \tag{361}$$

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 $E \longrightarrow 0$ again a free theory. Hence, the irrelevant perturbation in Eq. (361) 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, for 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 $\mathrm{W}^{\pm}$ boson:



At energies below $M_{\mathrm{W}} = 80.4$ GeV the dynamics of the W^+ boson is dominated by its mass term and therefore becomes non-propagating, giving rise to the effective four-fermion Fermi theory.

To summarize our discussion so far, we have 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

$$\mathcal{L}_0 = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\overline{\psi}D\!\!\!\!/\,\psi \tag{362}$$

is restored at low energies or not, we need to study the perturbed theory in more detail. This we did in Section 7.1 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 behaviour 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

$$\begin{aligned}\mathcal{L} &= -\frac{1}{4}\left(\partial_{\mu}A^{a}_{\nu} - \partial_{\nu}A^{a}_{\mu}\right)\left(\partial^{\mu}A^{a\,\nu} - \partial^{\nu}A^{a\,\mu}\right) - 4gf^{abc}A^{a}_{\mu}A^{b}_{\nu}\,\partial^{\mu}A^{c\,\nu} \\ &+ g^2 f^{abc}f^{ade}A^{b}_{\mu}A^{c}_{\nu}A^{d\,\mu}A^{e\,\nu} \equiv \mathcal{L}_0 + \mathcal{O}_g \; ,\end{aligned} \tag{363}$$

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 from the fixed point in this limit.

It is very important to note here that in the Wilsonian view the cut-off is not necessarily regarded as just some artefact 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 cut-off $\Lambda = M_{\rm W}$ at which the theory has to be replaced by the Standard Model. In the case of the Standard Model itself the cut-off can be taken at the 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 cut-off 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 cut-off, there is no fundamental difference between renormalizable and non-renormalizable theories. Actually, a renormalizable field theory, like the Standard Model, can generate non-renormalizable 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 cut-off, $E \ll \Lambda$, since their contribution to the amplitudes will be suppressed by powers of $\frac{E}{\Lambda}$.

8 Special topics

8.1 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:

$$\widehat{n} \equiv \int \frac{d^3k}{(2\pi)^3} \alpha^\dagger(\vec{k})\alpha(\vec{k}) \ , \qquad [\widehat{H}, \widehat{n}] = 0 \ . \tag{364}$$

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 archetypal 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

$$S = \int d^4x \left[\frac{1}{2}\partial_\mu\phi(x)\partial^\mu\phi(x) - \frac{m^2}{2}\phi(x)^2 + J(x)\phi(x)\right] \ , \tag{365}$$

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

$$\left(\nabla^2 + m^2\right)\phi(x) = J(x) \ . \tag{366}$$

Our plan is to quantize this theory but, unlike the case analysed 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

$$\phi(x) = \phi_0(x) + i\int d^4x'\, G_R(x - x')J(x') \ , \tag{367}$$

where $\phi_0(x)$ is a general solution to the homogeneous equation and

$$G_R(t,\vec{x}) = \int \frac{d^4k}{(2\pi)^4}\frac{i}{k^2-m^2}e^{-ik\cdot x} = i\,\theta(t)\int\frac{d^3k}{(2\pi)^3}\frac{1}{2\omega_k}\left(e^{-i\omega_k t+\vec{k}\cdot\vec{x}} - e^{i\omega_k t - i\vec{p}\cdot\vec{x}}\right), \tag{368}$$

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[15].

To make the argument simpler we assume that $J(x)$ is switched on at $t = 0$, and only lasts for a time τ, that is

$$J(t,\vec{x}) = 0 \qquad \text{if } t < 0 \text{ or } t > \tau\ . \tag{369}$$

We are interested in a solution of Eq. (366) for times after the external source has been switched off, $t > \tau$. In this case the expression (368) can be written in terms of the Fourier modes $\widetilde{J}(\omega,\vec{k})$ of the source as

$$\phi(t,\vec{x}) = \phi_0(x) + i\int\frac{\mathrm{d}^3k}{(2\pi)^3}\frac{1}{2\omega_k}\left[\widetilde{J}(\omega_k,\vec{k})e^{-i\omega_k t+i\vec{k}\cdot\vec{x}} - \widetilde{J}(\omega_k,\vec{k})^* e^{i\omega_k t - i\vec{k}\cdot\vec{x}}\right]\ . \tag{370}$$

On the other hand, the general solution $\phi_0(x)$ has already been computed in Eq. (77). Combining this result with Eq. (370) we find the following expression for the late-time general solution to the Klein–Gordon equation in the presence of the source:

$$\begin{aligned}\phi(t,x) &= \int\frac{\mathrm{d}^3k}{(2\pi)^3}\frac{1}{\sqrt{2\omega_k}}\left\{\left[\alpha(\vec{k}) + \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k})\right]e^{-i\omega_k t+i\vec{k}\cdot\vec{x}}\right.\\ &+ \left.\left[\alpha^*(\vec{k}) - \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k})^*\right]e^{i\omega_k t-i\vec{k}\cdot\vec{x}}\right\}\ .\end{aligned} \tag{371}$$

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. (367) and (368) 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 (371). 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

$$\begin{aligned}\alpha(\vec{k}) &\longrightarrow \beta(\vec{k}) \equiv \alpha(\vec{k}) + \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k})\ ,\\ \alpha^\dagger(\vec{k}) &\longrightarrow \beta^\dagger(\vec{k}) \equiv \alpha^\dagger(\vec{k}) - \frac{i}{\sqrt{2\omega_k}}\widetilde{J}(\omega_k,\vec{k})^*\ .\end{aligned} \tag{372}$$

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:

$$\left.\begin{aligned}\alpha(\vec{k})|0_-\rangle = 0\\ \beta(\vec{k})|0_+\rangle = 0\end{aligned}\right\} \qquad \forall\,\vec{k}\ . \tag{373}$$

[15] 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 $J(x)$.

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

$$\widehat{H}^{(-)} = \int d^3k\, \omega_k\, \alpha^\dagger(\vec{k})\alpha(\vec{k}) \tag{374}$$

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 these operators the Hamiltonian is written as

$$\widehat{H}^{(+)} = \int d^3k\, \omega_k\, \beta^\dagger(\vec{k})\beta(\vec{k})\ . \tag{375}$$

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 representation, 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,

$$\widehat{n}^{(-)}|0_-\rangle = 0\ . \tag{376}$$

However the late-time observer will find that the state $|0_-\rangle$ contains an average number of particles given by

$$\langle 0_-|\widehat{n}^{(+)}|0_-\rangle = \int \frac{d^3k}{(2\pi)^3}\frac{1}{2\omega_k}\left|\widetilde{J}(\omega_k,\vec{k})\right|^2\ . \tag{377}$$

Moreover, $|0_-\rangle$ is no longer the ground state for the 'out' observer. On the contrary, this state has a vacuum expectation value for $\widehat{H}^{(+)}$

$$\langle 0_-|\widehat{H}^{(+)}|0_-\rangle = \frac{1}{2}\int \frac{\mathrm{d}^3k}{(2\pi)^3}\left|\tilde{J}(\omega_k,\vec{k})\right|^2\ . \tag{378}$$

The key to understanding 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 state $|0_-\rangle$ is not the vacuum of his Hamiltonian, but actually an excited state that is a superposition of states with a 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. (378). 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 [40] 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 Wentzel–Kramers–Brillouin (WKB) approximation.

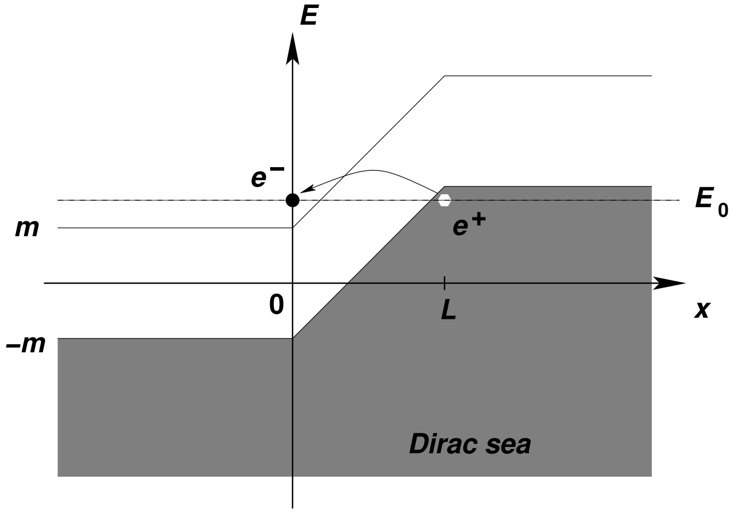


Fig. 16: Pair creation by an electric field in the Dirac sea picture

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 an electrostatic potential

$$V(\vec{r}) = \left\{ \begin{array}{cc} 0 & x < 0 \\ \mathcal{E}(x - x_0) & 0 < x < L \\ \mathcal{E}L & x > L \end{array} \right. . \tag{379}$$

After the field has been switched on, the Dirac sea looks like in Fig. 16. In particular we find that if $\mathcal{E}L > 2m$ 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[16]

$$\begin{aligned} T_{\rm WKB} &= \exp\left[-2\int_{\frac{1}{e\mathcal{E}}\left(E_0-\sqrt{m^2+\vec{p}_T^{\,2}}\right)}^{\frac{1}{e\mathcal{E}}\left(E_0+\sqrt{m^2+\vec{p}_T^{\,2}}\right)} dx\,\sqrt{m^2 - [E_0 - e\mathcal{E}(x-x_0)]^2 + \vec{p}_T^{\,2}}\right] \\ &= \exp\left[-\frac{\pi}{e\mathcal{E}}\left(\vec{p}_T^{\,2} + m^2\right)\right] , \end{aligned} \tag{380}$$

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

$$W = 2\left(\frac{e\mathcal{E}}{2\pi}\right)\int\frac{d^2p_T}{(2\pi)^2}e^{-\frac{\pi}{e\mathcal{E}}\left(\vec{p}_T^{\,2}+m^2\right)} = \frac{e^2\mathcal{E}^2}{4\pi^3}e^{-\frac{\pi\, m^2}{e\mathcal{E}}} , \tag{381}$$

[16] Note that the electron satisfies 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 $p_x = 0$.

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 Eq. (381):

$$W = \frac{e^2\mathcal{E}^2}{4\pi^3 c\hbar^2} e^{-\frac{\pi\, m^2 c^3}{\hbar e \mathcal{E}}} . \tag{382}$$

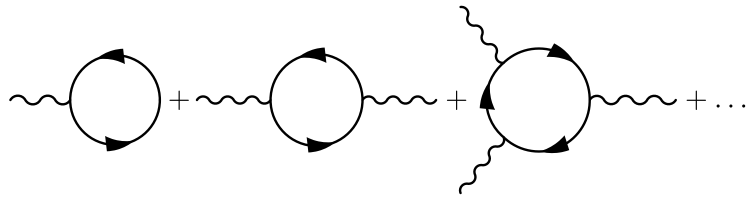
The exponential suppression of the pair production disappears when the electric field reaches the critical value $\mathcal{E}_{\text{crit}}$ at which the exponent is of order one:

$$\mathcal{E}_{\text{crit}} = \frac{m^2 c^3}{\hbar e} \simeq 1.3 \times 10^{16}\ \text{V cm}^{-1} . \tag{383}$$

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 [41] 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:

$$i\Gamma[A_\mu] \equiv$$



$$= \log\ \det\left[1 - ie\not{A}\frac{1}{i\not{\partial} - m}\right] . \tag{384}$$

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

$$W = \frac{e^2\mathcal{E}^2}{4\pi^3}\sum_{n=1}^{\infty}\frac{1}{n^2}e^{-n\frac{\pi\, m^2}{e\mathcal{E}}} . \tag{385}$$

Our simple argument based on tunnelling in the Dirac sea gave only the leading term of Schwinger's result (385). The remaining terms can also be captured in the WKB approximation by taking into account the probability of production of several pairs, i.e., the tunnelling 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 analysed above, implies that it is no longer possible 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 happens in a cosmological set-up or for a collapsing star) different observers will identify different vacuum states. As a consequence, what one observer calls the vacuum will be full of particles for a different observer. This is precisely what is behind the phenomenon of

Hawking radiation [42]. 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 contains particles for an observer in the asymptotic future. As a consequence, a detector located far away from the black hole detects a stream of thermal radiation with temperature

$$T_{\text{Hawking}} = \frac{\hbar c^3}{8\pi G_{\text{N}}\, k\, M} \tag{386}$$

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 results can be obtained. A more heuristic way is perhaps to think of this particle creation as resulting from quantum tunnelling of particles across the potential barrier posed by gravity [43].

8.2 Supersymmetry

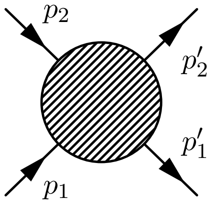
One of the things that we have learned in our journey around the landscape of quantum field theory is that our knowledge of the fundamental interactions in Nature is based on the idea of symmetry, and in particular gauge symmetry. The Lagrangian of the Standard Model can be written just including all possible renormalizable terms (i.e., with canonical dimension smaller than or 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.

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

$$\mathcal{A}(\text{in} \longrightarrow \text{out}) = \langle \vec{p}_1, \ldots ; \text{in}|\widehat{S}|\vec{p}_1{}', \ldots ; \text{out}\rangle \ . \tag{387}$$

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 shall 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

$$t = (p_1' - p_1)^2 = 2\left(m_1^2 - p_1 \cdot p_1'\right) = 2\left(m_1^2 - E_1 E_1' + |\vec{p}_1||\vec{p}_1{}'|\cos\vartheta\right) \ . \tag{388}$$

If there were 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 the analyticity of the scattering amplitudes is to introduce 'fermionic' symmetries, i.e., symmetries whose generators are anticommuting objects [44]. This means that in addition to the generators of the Poincaré group[17] P^μ,

[17] The generators $M^{\mu\nu}$ are related with the ones for boost and rotations introduced in Section 4.1 by $J^i \equiv M^{0i}$, $M^i = \frac{1}{2}\varepsilon^{ijk}M^{jk}$. In this section we also use the 'dotted spinor' notation, in which spinors in the $(\frac{1}{2}, \mathbf{0})$ and $(\mathbf{0}, \frac{1}{2})$ representations of the Lorentz group are indicated by undotted $(a, b, \ldots)$ and dotted $(\dot{a}, \dot{b}, \ldots)$ indices, respectively.

$M^{\mu\nu}$ 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 [45]

$$\begin{aligned} \{Q^I_a, \overline{Q}_{\dot{b}\,J}\} &= 2\sigma^\mu_{a\dot{b}} P_\mu \delta^I{}_J \,, \\ \{Q^I_a, Q^J_b\} &= 2\varepsilon_{ab}\mathcal{Z}^{IJ} \,, & (389) \\ \{\overline{Q}^I_{\dot{a}}, \overline{Q}^J_{\dot{b}}\} &= -2\varepsilon_{\dot{a}\dot{b}}\overline{\mathcal{Z}}^{IJ} \,, & (390) \end{aligned}$$

where $\mathcal{Z}^{IJ} \in \mathbb{C}$ commutes 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, $Q_{\dot{a}\,J}$:

$$\begin{aligned} [Q^I_a, P^\mu] &= [\overline{Q}_{\dot{a}\,I}, P^\mu] = 0 \,, \\ [Q^I_a, M^{\mu\nu}] &= \frac{1}{2}(\sigma^{\mu\nu})_a{}^b Q^I_b \,, & (391) \\ [\overline{Q}_{a\,I}, M^{\mu\nu}] &= -\frac{1}{2}(\overline{\sigma}^{\mu\nu})_{\dot{a}}{}^{\dot{b}} \overline{Q}_{\dot{b}\,I} \,, \end{aligned}$$

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 in the $(\frac{\mathbf{1}}{\mathbf{2}}, \mathbf{0})$ and $(\mathbf{0}, \frac{\mathbf{1}}{\mathbf{2}})$ representations of the Lorentz group, respectively.

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 we start with the case $\mathcal{N} = 1$ in which there is a single pair of supercharges Q_a, $\overline{Q}_{\dot{a}}$ satisfying the algebra

$$\{Q_a, \overline{Q}_{\dot{b}}\} = 2\sigma^\mu_{a\dot{b}} P_\mu \,, \qquad \{Q_a, Q_b\} = \{\overline{Q}_{\dot{a}}, \overline{Q}_{\dot{b}}\} = 0 \,. \tag{392}$$

Note that in the $\mathcal{N} = 1$ case there is no possibility of having central charges.

We study now the representations of the supersymmetry algebra (392), 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. (392) is given by

$$2\sigma^\mu_{a\dot{b}} P_\mu = 2(P^0 - \sigma^3 P^3) = \begin{pmatrix} 0 & 0 \\ 0 & 4E \end{pmatrix} \,. \tag{393}$$

Therefore the algebra of supercharges in the massless case reduces to

$$\begin{aligned} \{Q_1, Q_1^\dagger\} &= \{Q_1, Q_2^\dagger\} = 0 \,, \\ \{Q_2, Q_2^\dagger\} &= 4E \,. & (394) \end{aligned}$$

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 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

$$a = \frac{1}{2\sqrt{E}} Q_2 \,, \qquad a^\dagger = \frac{1}{2\sqrt{E}} Q_2^\dagger \,, \tag{395}$$

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

$$|\lambda\rangle \,, \qquad |\lambda + \tfrac{1}{2}\rangle \equiv a^\dagger|\lambda\rangle \,. \tag{396}$$

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

$$|0^+\rangle\,, \qquad |\tfrac{1}{2}\rangle \equiv a^\dagger|0^+\rangle\,. \tag{397}$$

  However, this multiplet is not invariant under the CPT transformation which reverses the sign of the helicity of the states. In order to have a CPT-invariant theory we have to add to this multiplet its CPT conjugate which can be obtained from a vacuum state with helicity $\lambda = -\frac{1}{2}$

$$|0^-\rangle\,, \qquad |-\tfrac{1}{2}\rangle\,. \tag{398}$$

  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 $\lambda = 1$

$$|\tfrac{1}{2}\rangle\,, \qquad |1\rangle \equiv a^\dagger|\tfrac{1}{2}\rangle. \tag{399}$$

  As with the scalar multiplet we add the CPT conjugate obtained from a vacuum state with helicity $\lambda = -1$

$$|-\tfrac{1}{2}\rangle\,, \qquad |-1\rangle\,, \tag{400}$$

  which together with Eq. (399) 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.

$$\begin{array}{rll}
\lambda = 2: & a_1^\dagger \ldots a_8^\dagger|-2\rangle & \binom{8}{8} = 1 \text{ state}\,, \\
\lambda = \dfrac{3}{2}: & a_{I_1}^\dagger \ldots a_{I_7}^\dagger|-2\rangle & \binom{8}{7} = 8 \text{ states}\,, \\
\lambda = 1: & a_{I_1}^\dagger \ldots a_{I_6}^\dagger|-2\rangle & \binom{8}{6} = 28 \text{ states}\,, \\
\lambda = \dfrac{1}{2}: & a_{I_1}^\dagger \ldots a_{I_5}^\dagger|-2\rangle & \binom{8}{5} = 56 \text{ states}\,, \\
\lambda = 0: & a_{I_1}^\dagger \ldots a_{I_4}^\dagger|-2\rangle & \binom{8}{4} = 70 \text{ states}\,, \\
\lambda = -\dfrac{1}{2}: & a_{I_1}^\dagger a_{I_2}^\dagger a_{I_3}^\dagger|-2\rangle & \binom{8}{3} = 56 \text{ states}\,, \\
\lambda = -1: & a_{I_1}^\dagger a_{I_2}^\dagger|-2\rangle & \binom{8}{2} = 28 \text{ states}\,, \\
\lambda = -\dfrac{3}{2}: & a_{I_1}^\dagger|-2\rangle & \binom{8}{1} = 8 \text{ states}\,, \\
\lambda = -2: & |-2\rangle & \qquad 1 \text{ state}\,.
\end{array} \tag{401}$$

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:

$$\begin{array}{lrl} \text{bosons:} & 1+28+70+28+1=128 & \text{states ,} \\ \text{fermions:} & 8+56+56+8=128 & \text{states .} \end{array}$$

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

$$\{Q^I_\alpha, \overline{Q}_{\dot{\beta}J}\} = 2M\delta_{\alpha\dot{\beta}}\delta^I{}_J \; . \tag{402}$$

We proceed now in a similar way to the massless case by defining the operators

$$a^I_\alpha \equiv \frac{1}{\sqrt{2M}}Q^I_\alpha \; , \qquad a^\dagger_{\dot{\alpha}I} \equiv \frac{1}{\sqrt{2M}}\overline{Q}_{\dot{\alpha}I} \; . \tag{403}$$

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:

$$|0\rangle \; , \qquad a^\dagger_{\dot{\alpha}}|0\rangle, \qquad \varepsilon^{\dot{\alpha}\dot{\beta}}a^\dagger_{\dot{\alpha}}a^\dagger_{\dot{\beta}}|0\rangle \; , \tag{404}$$

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 as[18]

$$E_{\text{vac}} = \pm\frac{1}{2}\delta(\vec{0})\int d^3p\,\omega_p \; , \tag{405}$$

where the sign ± corresponds to bosons and fermions, respectively. 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.

Acknowledgements

It is a great pleasure to thank the organizers of the school, and in particular Teresa Dova, for the opportunity to present this material, and for the wonderful atmosphere they created throughout the school. The work of Miguel A. Vázquez-Mozo has been partially supported by Spanish Science Ministry Grants FPA2002-02037, FPA2005-04823 and BFM2003-02121.

[18] 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.

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 $D(g)$

$$D(g) : V \longrightarrow V \tag{A.1}$$

satisfying the group operations

$$D(g_1)D(g_2) = D(g_1 g_2) , \qquad D(g_1^{-1}) = D(g_1)^{-1} , \qquad g_1, g_2 \in \mathcal{G} . \tag{A.2}$$

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

$$[D(g), A] = 0, \ \forall g \qquad \Longleftrightarrow \qquad A = \lambda \mathbf{1} , \quad \lambda \in \mathbb{C} . \tag{A.3}$$

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 especially 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 shall 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 neighbourhood of the identity element they can be expressed in terms of a set of generators T^a ($a = 1, \ldots, \dim G$) as

$$D(g) = \exp[-i\alpha_a T^a] \equiv \sum_{n=0}^{\infty} \frac{(-i)^n}{n!} \alpha_{a_1} \ldots \alpha_{a_n} T^{a_1} \ldots T^{a_n} , \tag{A.4}$$

where $\alpha_a \in \mathbb{C}$ are a set of coordinates of $\mathcal{M}$ in a neighbourhood 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, which in general has the form

$$[T^a, T^b] = i f^{abc} T^c , \tag{A.5}$$

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 concepts 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

$$D_{\frac{1}{2}}(\alpha_i) = e^{-\frac{i}{2}\alpha_i \sigma^i} , \tag{A.6}$$

with σ^i the Pauli matrices; the other is the adjoint (or spin 1) representation, which can be written as

$$D_1(\alpha_i) = e^{-i\alpha_i J^i} , \tag{A.7}$$

where

$$J^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} , \qquad J^2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} , \qquad J^3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} . \tag{A.8}$$

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 also be constructed with dimension

$$\dim D_j(g) = 2j + 1 \, . \tag{A.9}$$

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 colour gauge symmetry SU(3). The elements of these representations can be written as

$$D_{\mathbf{3}}(\alpha^a) = e^{\frac{i}{2}\alpha^a \lambda_a} \, , \qquad D_{\overline{\mathbf{3}}}(\alpha^a) = e^{-\frac{i}{2}\alpha^a \lambda_a^T} \qquad (a = 1, \ldots, 8) \, , \tag{A.10}$$

where λ_a are the eight Hermitian Gell-Mann matrices

$$\begin{aligned}
\lambda_1 &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} , \qquad \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} , \qquad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
\lambda_4 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} , \qquad \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix} , \qquad \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} , \\
\lambda_7 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} , \qquad \lambda_8 = \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & -\frac{2}{\sqrt{3}} \end{pmatrix} .
\end{aligned} \tag{A.11}$$

Hence the generators of the representations $\mathbf{3}$ and $\overline{\mathbf{3}}$ are given by

$$T^a(\mathbf{3}) = \frac{1}{2}\lambda_a \, , \qquad T^a(\overline{\mathbf{3}}) = -\frac{1}{2}\lambda_a^T \, . \tag{A.12}$$

Irreducible representations can be classified in three groups: real, pseudoreal, and complex.

- 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

  $$\overline{T}^a = -ST^aS^{-1} \, , \qquad S^T = S \, . \tag{A.13}$$

  This is, for example, the case for the adjoint representation of SU(2) generated by the matrices (A.8)
- Pseudoreal representations: These are the representations for which an *antisymmetric matrix* S exists with the property

  $$\overline{T}^a = -ST^aS^{-1} \, , \qquad S^T = -S \, . \tag{A.14}$$

  As an example we can mention the spin-$\frac{1}{2}$ representation of SU(2) generated by $\frac{1}{2}\sigma^i$.
- Complex representations: A representation is complex if the generators and their complex conjugate are not related by a similarity transformation. This is for instance the case for 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^a_R are the generators in a certain representation R of the Lie algebra, it is easy to see that the matrix $\sum_{a=1}^{\dim G} T^a_R T^a_R$ commutes with every generator T^a_R. Therefore, because of Schur's lemma, it has to be proportional to the identity[19]. This defines the Casimir invariant $C_2(R)$ as

$$\sum_{a=1}^{\dim G} T^a_R T^a_R = C_2(R)\mathbf{1}. \tag{A.15}$$

A second invariant $T_2(R)$ associated with a representation R can also be defined by the identity

$$\mathrm{Tr}\, T^a_R T^b_R = T_2(R)\delta^{ab}\ . \tag{A.16}$$

Actually, taking the trace in Eq. (A.15) and combining the result with Eq. (A.16) we find that both invariants are related by the identity

$$C_2(R)\dim R = T_2(R)\dim G\ , \tag{A.17}$$

with $\dim R$ the dimension of the representation R.

These two invariants appear frequently in quantum field theory calculations with non-Abelian 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

$$C_2(\text{fund}) = \frac{N^2-1}{2N}\ , \qquad T_2(\text{fund}) = \frac{1}{2}\ , \tag{A.18}$$

whereas for the adjoint representation the results are

$$C_2(\text{adj}) = N\ , \qquad T_2(\text{adj}) = N\ . \tag{A.19}$$

A third invariant $A(R)$ is especially important in the calculation of anomalies. As discussed in Section 6, the chiral anomaly in gauge theories is proportional to the group theoretical factor Tr $[T^a_R\{T^b_R, T^c_R\}]$. This leads us to define $A(R)$ as

$$\mathrm{Tr}\ \Big[T^a_R\{T^b_R, T^c_R\}\Big] = A(R), d^{abc} \tag{A.20}$$

where d^{abc} is symmetric in its three indices and does not depend on the representation. Therefore the cancellation of anomalies in a gauge theory with fermions transformed in the representation R of the gauge group is guaranteed if the corresponding invariant $A(R)$ vanishes.

It is not difficult to prove that $A(R) = 0$ if the representation R is either real or pseudoreal. Indeed, if this is the case, then there is a matrix S (symmetric or antisymmetric) that intertwines the generators T^a_R and their complex conjugates $\overline{T}^a_R = -ST^a_R S^{-1}$. Then, using the hermiticity of the generators we can write

$$\mathrm{Tr}\ \Big[T^a_R\{T^b_R, T^c_R\}\Big] = \mathrm{Tr}\ \Big[T^a_R\{T^b_R, T^c_R\}\Big]^T = \mathrm{Tr}\ \Big[\overline{T}^a_R\{\overline{T}^b_R, \overline{T}^c_R\}\Big]\ . \tag{A.21}$$

Now, using Eq. (A.13) or Eq. (A.14) we have

$$\mathrm{Tr}\ \Big[\overline{T}^a_R\{\overline{T}^b_R, \overline{T}^c_R\}\Big] = -\mathrm{Tr}\ \Big[ST^a_R S^{-1}\{ST^b_R S^{-1}, ST^c_R S^{-1}\}\Big] = -\mathrm{Tr}\ \Big[T^a_R\{T^b_R, T^c_R\}\Big]\ , \tag{A.22}$$

which proves that Tr $\big[T^a_R\{T^b_R, T^c_R\}\big]$ 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.

[19] Schur's lemma states that a representation of a group is irreducible if and only if all matrices commuting with every element of the representation are proportional to the identity.

References

[1] J.D. Bjorken and S.D. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965).

[2] C. Itzykson and J.-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980).

[3] P. Ramond, *Field Theory: A Modern Primer*, 2nd ed. (Addison-Wesley, Redwood City, CA, 1989).

[4] M.E. Peskin and D.V. Schroeder, *An Introduction to Quantum Field Theory* (Addison-Wesley, Reading, MA, 1995).

[5] S. Weinberg, *The Quantum Theory of Fields*, 3 vols. (Cambridge University Press, Cambridge, 1995).

[6] P. Deligne *et al.* (editors), *Quantum Fields and Strings: A Course for Mathematicians*, 2 vols. (American Mathematical Society, Providence, RI, 1999–2000).

[7] A. Zee, *Quantum Field Theory in a Nutshell* (Princeton University, Princeton, 2003).

[8] B.S. DeWitt, *The Global Approach to Quantum Field Theory*, 2 vols. (Clarendon Press, Oxford, 2003).

[9] V.P. Nair, *Quantum Field Theory: A Modern Perspective* (Springer, 2005).

[10] O. Klein, Die Reflexion von Elektronen an einem Potentialsprung nach der Relativischen Dynamik von Dirac, *Z. Phys.* **53** (1929) 157.

[11] B.R. Holstein, Klein's paradox, *Am. J. Phys.* **66** (1998) 507.

[12] N. Dombey and A. Calogeracos, Seventy years of the Klein paradox, *Phys. Rep.* **315** (1999) 41.
N. Dombey and A. Calogeracos, History and physics of the Klein paradox, *Contemp. Phys.* **40** (1999) 313 [quant-ph/9905076].

[13] F. Sauter, Zum Kleinschen Paradoxon, *Z. Phys.* **73** (1932) 547.

[14] H.B.G. Casimir, On the attraction between two perfectly conducting plates, *Proc. Kon. Ned. Akad. Wet.* **60** (1948) 793.

[15] G. Plunien, B. Müller and W. Greiner, The Casimir effect, *Phys. Rep.* **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.

[16] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972).

[17] M.J. Sparnaay, Measurement of attractive forces between flat plates, *Physica* **24** (1958) 751.

[18] Y. Aharonov and D. Bohm, Significance of the electromagnetic potentials in the quantum theory, *Phys. Rev.* **115** (1959) 485.

[19] P.A.M. Dirac, Quantised singularities in the electromagnetic field, *Proc. R. Soc.* **133** (1931) 60.

[20] P.A.M. Dirac, *Lectures on Quantum Mechanics* (Dover, 2001).

[21] M. Henneaux and C. Teitelboim, *Quantization of Gauge Systems* (Princeton University, Princeton, 1992).

[22] 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.

[23] P. Ramond, *Journeys Beyond the Standard Model* (Perseus Books, Cambridge, MA, 1999).
R. N. Mohapatra, *Unification and Supersymmetry: The Frontiers of Quark–Lepton Physics*, 3rd ed. (Springer, Berlin, 2003).

[24] C.P. Burguess, Goldstone and pseudogoldstone bosons in nuclear, particle and condensed matter physics, *Phys. Rep.* **330** (2000) 193 [hep-th/9808176].

[25] L. Álvarez-Gaumé, An introduction to anomalies, in *Fundamental Problems of Gauge Field Theory,* Eds. G. Velo and A.S. Wightman (Plenum, New York, 1986).

[26] R. Jackiw, Topological investigations of quantized gauge theories, in *Current Algebra and Anomalies,* Eds. S.B. Treiman, R.W. Jackiw, E. Witten and B. Zumino (World Scientific, Singapore, 1986).

[27] 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 Cim.* **A60** (1969) 47.

[28] F.J. Ynduráin, *The Theory of Quark and Gluon Interactions*, 3rd ed. (Springer, Berlin, 1999).

[29] G. 't Hooft, How the instantons solve the U(1) problem, *Phys. Rep.* **142** (1986) 357.

[30] 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.

[31] J. Steinberger, On the use of subtraction fields and the lifetimes of some types of meson decay, *Phys. Rev.* **76** (1949) 1180.

[32] S.L. Adler and W.A. Bardeen, Absence of higher order corrections in the anomalous axial vector divergence equation, *Phys. Rev.* **182** (1969) 1517.

[33] E. Witten, An SU(2) anomaly, *Phys. Lett.* **B117** (1982) 324.

[34] D.J. Gross and F. Wilczek, Ultraviolet behavior of nonabelian gauge theories, *Phys. Rev. Lett.* **30** (1973) 1343.

[35] H.D. Politzer, Reliable perturbative results for strong interactions?, *Phys. Rev. Lett.* **30** (1973) 1346.

[36] G. 't Hooft, remarks at the *Colloquium on Renormalization of Yang–Mills Fields and Applications to Particle Physics*, Marseille, 1972 (CNRS, Marseille, 1972).

[37] 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].

[38] 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.

[39] L.P. Kadanoff, Scaling laws for Ising models near T_c, *Physics* **2** (1966) 263.

[40] J. Schwinger, On gauge invariance and vacuum polarization, *Phys. Rev.* **82** (1951) 664.

[41] E. Brezin and C. Itzykson, Pair production in vacuum by an alternating field, *Phys. Rev.* **D2** (1970) 1191.

[42] S.W. Hawking, Particle creation by black holes, *Commun. Math. Phys.* **43** (1975) 199.

[43] M.K. Parikh and F. Wilczek, Hawking radiation as tunneling, *Phys. Rev. Lett.* **85** (2000) 5042 (hep-th/9907001).

[44] 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.

[45] R. Haag, J. Łopuszański and M. Sohnius, All possible generators of supersymmetries of the S-matrix, *Nucl. Phys.* **B88** (1975) 257.

Introduction to QCD in hadronic collisions

M.L. Mangano
CERN, Geneva, Switzerland

Abstract
I review in this series of lectures the basics of perturbative quantum chromodynamics and some simple applications to the physics of high-energy hadronic collisions.

1 Introduction

Quantum chromodynamics (QCD) is the theory of strong interactions. It is formulated in terms of elementary fields (quarks and gluons), whose interactions obey the principles of a relativistic quantum field theory, with a non-Abelian gauge invariance SU(3). The emergence of QCD as a theory of strong interactions could be reviewed historically, analysing the various experimental data and the theoretical ideas available in the years 1960–1973 (see, for example, Refs. [17, 18]). To do this accurately and usefully would require more time than I have available. I therefore prefer to introduce QCD right away, and to use my time in exploring some of its consequences and applications. I will therefore assume that you all know more or less what QCD is! I assume you know that hadrons are made of quarks, that quarks are spin-1/2, colour-triplet fermions, interacting via the exchange of an octet of spin-1 gluons. I assume you know the concept of running couplings, asymptotic freedom and of confinement. I shall finally assume that you have some familiarity with the fundamental ideas and formalism of quantum electrodynamics (QED): Feynman rules, renormalization, gauge invariance.

If you go through lecture series on QCD (e.g., the lectures given in previous years at the European Schools of High-Energy Physics, Refs. [9–11]), you will hardly ever find the same item twice. This is because QCD today covers a huge set of subjects and each of us has his own concept of what to do with QCD and of what are the 'fundamental' notions of QCD and its 'fundamental' applications. As a result, you will find lecture series centred around non-perturbative applications, (lattice QCD, sum rules, chiral perturbation theory, heavy-quark effective theory), around formal properties of the perturbative expansion (asymptotic behaviour, renormalons), techniques to evaluate complex classes of Feynman diagrams, or phenomenological applications of QCD to possibly very different sets of experimental data: structure functions, deep-inelastic scattering (DIS) sum rules, polarized DIS, small-x physics (including hard pomerons, diffraction), LEP physics, $p\bar{p}$ collisions, etc.

I will not be able to cover or even to mention all of this. After introducing some basic material, I will focus on some elementary applications of QCD in high-energy phenomena, and in particular on the case of hadron–hadron collisions. The material covered in these lectures includes the following.

1. Gauge invariance and Feynman rules for QCD.
2. The structure of the proton.
3. The evolution of final states: from quarks and gluons to hadrons.
4. Some key hard processes in hadron–hadron collisions: formalism, W/Z production, jet production.

The treatment will be very elementary, and the emphasis will be on basic and intuitive physics concepts. Explicit details and the derivation of equations and formulas is left for a few appendices, covering

a. renormalization, running coupling, renormalization group invariance;
b. deep-inelastic scattering and evolution equations;
c. jet observables in e^+e^- collisions.

Given the large number of papers which have contributed to the development of the field, it is impossible to provide a fair bibliography. I therefore limit my list of references to some excellent books and review articles covering the material presented here, and more. Papers on specific items can be easily found by consulting the standard `hep-th` and `hep-ph` preprint archives.

2 QCD Feynman rules

Before starting with the applications, we need to spend some time developing the formalism and the necessary theoretical ideas. I will start from the Feynman rules. I will use an approach which is not canonical, namely it does not follow the standard path of the construction of a gauge-invariant Lagrangian and the derivation of Feynman rules from it. I will rather start from QED, and empirically construct the extension to a non-Abelian theory by enforcing the desired symmetries directly on some specific scattering amplitudes. Hopefully, this will lead to a better insight into the relation between gauge invariance and Feynman rules. It will also provide you with a way of easily recalling or checking your rules when books are not around!

2.1 Summary of QED Feynman rules

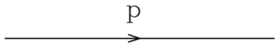
We start by summarizing the familiar Feynman rules for QED. They are obtained from the Lagrangian

$$\mathcal{L} \;=\; \bar{\psi}(i\not\partial - m)\psi - e\bar{\psi}\not A\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \qquad (1)$$

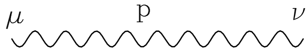
where ψ is the electron field, of mass m and coupling constant e, and $F^{\mu\nu}$ is the electromagnetic field strength.

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \; . \qquad (2)$$

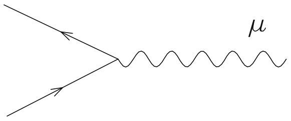
The resulting Feynman rules are summarized in the following table:



$$= \frac{i}{\not p - m + i\epsilon} = i\frac{\not p + m}{p^2 - m^2 + i\epsilon} \qquad (3)$$

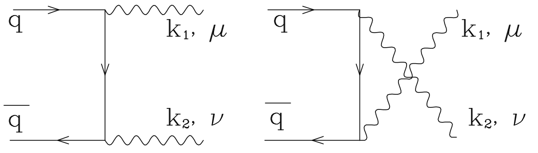


$$= -i\frac{g_{\mu\nu}}{p^2 + i\epsilon} \text{ (Feynman gauge)} \qquad (4)$$



$$= -ie\gamma_\mu Q \quad (Q = -1 \text{ for the electron, } Q = 2/3 \text{ for the } u\text{-quark, etc.}) \qquad (5)$$

Let us start by considering a simple QED process, $e^+e^- \to \gamma\gamma$ (for simplicity we shall always assume $m = 0$):



$$= D_1 + D_2 \; . \qquad (6)$$

The total amplitude M_γ is given by

$$\frac{i}{e^2}\, M_\gamma \;\equiv\; D_1 + D_2 = \bar{v}(\bar{q})\, \not\epsilon_2 \frac{1}{\not q - \not k_1} \not\epsilon_1\, u(q) \;+\; \bar{v}(\bar{q})\not\epsilon_1 \frac{1}{\not q - \not k_2} \not\epsilon_2\, n(q) \equiv M_{\mu\nu}\epsilon_1^\mu \epsilon_2^\nu \; . \qquad (7)$$

Gauge invariance demands that

$$\epsilon_2^\nu \partial^\mu M_{\mu\nu} = \epsilon_1^\mu \partial^\nu M_{\mu\nu} = 0 \; . \qquad (8)$$

$M_\mu \equiv M_{\mu\nu}\epsilon_2^\nu$ is in fact the current that couples to the photon k_1. Charge conservation requires $\partial_\mu M^\mu = 0$:

$$\begin{aligned}\partial_\mu M^\mu = 0 \;\Rightarrow\; & \frac{d}{dt}\int M^0 d^3x = \int \partial_0 M^0\, d^3x \\ = \; & \int \vec{\nabla}\cdot\vec{M}\, d^3x = \int_{S\to\infty} \vec{M}\cdot d\vec{\Sigma} = 0\;. \end{aligned} \tag{9}$$

In momentum space, this means

$$k_1^\mu M_\mu = 0\;. \tag{10}$$

Another way of saying this is that the theory is invariant if $\epsilon_\mu(k) \to \epsilon_\mu(k) + f(k)\, k_\mu$. This is the standard Abelian gauge invariance associated with the vector potential transformations:

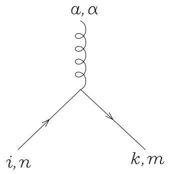
$$A_\mu(x) \to A_\mu(x) + \partial_\mu f(x)\;. \tag{11}$$

Let us verify that M_γ is indeed gauge invariant. Using $\not{q}u(q) = \bar{v}(\bar{q})\overline{\not{q}} = 0$ from the Dirac equation, we can rewrite $k_1^\mu M_\mu$ as

$$\begin{aligned} k_1^\mu \epsilon_2^\nu M_{\mu\nu} &= \bar{v}(\bar{q})\not{\epsilon}_2 \frac{1}{\not{q} - \not{k}_1}(\not{k}_1 - \not{q})u(q) \;+\; \bar{v}(\bar{q})(\not{k}_1 - \bar{q})\,\frac{1}{k_1 - \bar{q}}\not{\epsilon}_2 u(q) \\ &= -\bar{v}(\bar{q})\not{\epsilon}_2 u(q) + \bar{v}(\bar{q})\not{\epsilon}_2 u(q) = 0\;. \end{aligned} \tag{12}$$

Notice that the two diagrams are not individually gauge invariant, only the sum is. Notice also that the cancellation takes place independently of the choice of ϵ_2. The amplitude is therefore gauge invariant even in the case of emission of non-transverse photons.

Let us try now to generalize our QED example to a theory where the 'electrons' carry a non-Abelian charge, i.e., they transform under a non-trivial representation R of a non-Abelian group G (which, for the sake of simplicity, we shall always assume to be of the SU(N) type. Likewise, we shall refer to the non-Abelian charge as 'colour'). The standard current operator belongs to the product $R \otimes \bar{R}$. The only representation that belongs to $R \otimes \bar{R}$ for any R is the adjoint representation. Therefore the field that couples to the colour current must transform as the adjoint representation of the group G. So the only generalization of the photon field to the case of a non-Abelian symmetry is a set of vector fields transforming under the adjoint of G, and the simplest generalization of the coupling to fermions takes the form



$$= ig\lambda^a_{ki}\,\gamma^\mu_{mn} \tag{13}$$

where the matrices λ^a represent the algebra of the group on the representation R. By definition, they satisfy the algebra

$$[\lambda^a, \lambda^b] = if^{abc}\lambda^c \tag{14}$$

for a fixed set of structure constants f^{abc}, which uniquely characterize the algebra. We shall call quarks (q) the fermion fields in R and gluons (g) the vector fields which couple to the quark colour current.

The non-Abelian generalization of the $e^+e^- \to \gamma\gamma$ process is the $q\bar{q} \to gg$ annihilation. Its amplitude can be evaluated by including the λ matrices in Eq. (6):

$$\frac{i}{e^2}\, M_\gamma \to \frac{i}{g^2} M_g \equiv (\lambda^b\lambda^a)_{ij}\, D_1 + (\lambda^a\lambda^b)_{ij}\, D_2 \tag{15}$$

where (a, b) are colour labels (i.e., group indices) of gluons 1 and 2, and (i, j) are colour labels of $\bar{q}, q$, respectively. Using Eq. (14), we can rewrite Eq. (15) as

$$M_g = (\lambda^a \lambda^b)_{ij}\, M_\gamma - g^2\, f^{abc} \lambda^c_{ij}\, D_1 \; . \tag{16}$$

If we want the charge associated with the group G to be conserved, we still need to demand

$$k_1^\mu \epsilon_2^\nu \, M_g^{\mu\nu} = \epsilon_1^\mu k_2^\nu \, M_g^{\mu\nu} \; = \; 0 \; . \tag{17}$$

Substituting $\epsilon_1^\mu \to k_1^\mu$ in Eq. (16) we get instead, using Eq. (12),

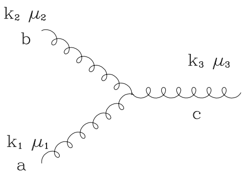
$$k_{1\,\mu} M_g^\mu = -g^2 f^{abc}\; \lambda^c_{ij}\; \bar{v}_i(\bar{q})\; \not{\epsilon}_2 u_i(q) \; . \tag{18}$$

The gauge cancellation taking place in QED between the two diagrams is spoiled by the non-Abelian nature of the coupling of quarks to gluons (i.e., λ^a and λ^b do not commute, and $f^{abc} \neq 0$).

The only possible way to solve this problem is to include additional diagrams. That new interactions should exist is by itself a reasonable fact, since gluons are charged (i.e., they transform under the symmetry group) and might want to interact among themselves. If we rewrite Eq. (18) as

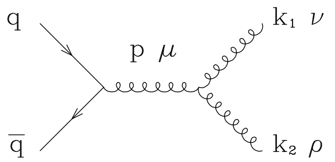
$$k_{1\,\mu} M_g^\mu \;=\; i \left(f^{abc}\, g\, \epsilon_2^\mu \right) \times \left(i\, g\, \lambda^c_{ij}\, \bar{v}(\bar{q})\, \gamma_\mu\, u(q) \right) \; , \tag{19}$$

we can recognize in the second factor the structure of the $q\bar{q}g$ vertex. The first factor has the appropriate colour structure to describe a triple-gluon vertex, with a, b, c the colour labels of the three gluons:



$$= \; g\, f^{abc}\, V_{\mu_1\mu_2\mu_3}(k_1, k_2, k_3) \; . \tag{20}$$

Equation (19) therefore suggests the existence of a coupling like Eq. (20), with a Lorentz structure $V_{\mu_1\mu_2\mu_3}$ to be specified, giving rise to the following contribution to $q\bar{q} \to gg$:



$$= \; -i\, g^2\, D_3 \;\; = \;\; (ig\, \lambda^a_{ij}) \bar{v}(\bar{q})_i\, \gamma^\mu\, u(q)_j \left(\frac{-i}{p^2} \right) \\ g\, f^{abc}\, V_{\mu\nu\rho}(-p, k_1, k_2)\, \epsilon_1^\nu(k_1)\, \epsilon_2^\rho(k_2) \; . \tag{21}$$

We now need to find $V_{\mu_1\mu_2\mu_3}(p_1, p_2, p_3)$ and to verify that the contribution of the new diagram to $k_1 \cdot M_g$ cancels that of the first two diagrams. We will now show that the constraints of Lorentz invariance, Bose symmetry and dimensional analysis uniquely fix V, up to an overall constant factor.

Dimensional analysis fixes the coupling to be linear in the gluon momenta. This is because each vector field carries dimension 1, there are three of them, and the interaction must have total dimension equal to 4. So at most one derivative (i.e., one power of momentum) can appear at the vertex. In principle, if some mass parameter were available, higher derivatives could be included, with the appropriate powers of the mass parameter appearing in the denominator. This is however not the case. It is important to remark that the absence of interactions with a higher number of derivatives is also crucial for the renormalizability of the interaction.

Lorentz invariance requires then that V be built out of terms of the form $g_{\mu_1\mu_2} p_{\mu_3}$. Bose symmetry requires V to be fully antisymmetric under the exchange of any pair $(\mu_i, p_i) \leftrightarrow (\mu_j, p_j)$ since the colour structure f^{abc} is totally antisymmetric. As a result, for example, a term like $g_{\mu_1\mu_2} p_3^{\mu_3}$ vanishes under antisymmetrization, while $g_{\mu_1\mu_2}p_1^{\mu_3}$ does not. Starting from this last term, we can easily add the pieces required to obtain the full antisymmetry in all three indices. The result is unique, up to an overall factor:

$$V_{\mu_1\mu_2\mu_3} \;=\; V_0\left[(k_1-k_2)^{\mu_3}g_{\mu_1\mu_2}+(k_2-k_3)^{\mu_1}g_{\mu_2\mu_3}+(k_3-k_1)^{\mu_2}g_{\mu_1\mu_3}\right] . \tag{22}$$

To test the gauge variation of the contribution D_3, we set $\mu_3=\mu, \epsilon_1=k_1$ and $k_3=-(k_1+k_2)$ in Eq. (21), and we get

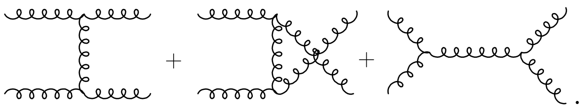
$$k_1^{\mu_1}\epsilon_2^{\mu_2}\,V_{\mu_1\mu_2\mu}(k_1,k_2,k_3) \;=\; V_0\left\{-(k_1+k_2)^{\mu}(k_1\cdot\epsilon_2)+2(k_1\cdot k_2)\epsilon_2^{\mu}-(k_2\cdot\epsilon_2)k_1^{\mu}\right\} . \tag{23}$$

The gauge variation is therefore

$$k_1\cdot D_3 = g^2 f^{abc}\lambda^c V_0\left[\bar{v}(\bar{q})\not{\epsilon}_2 u(q)-\frac{k_2\cdot\epsilon_2}{2k_1k_2}\bar{v}(\bar{q})\not{k}_1 u(q)\right] . \tag{24}$$

The first term cancels the gauge variation of D_1+D_2 provided $V_0=1$; the second term vanishes for a physical gluon k_2, since in this case $k_2\cdot\epsilon_2=0$. $D_1+D_2+D_3$ is therefore gauge invariant but, contrary to the case of QED, only for physical external on-shell gluons.

Having introduced a three-gluon coupling, we can induce processes involving only gluons, such as $gg\to gg$:



(25)

Once more it is necessary to verify the gauge invariance of this amplitude. It turns out that one more diagram is required, induced by a four-gluon vertex. Lorentz invariance, Bose symmetry and dimensional analysis uniquely determine once again the structure of this vertex. The overall factor is fixed by gauge invariance. The resulting Feynman rule for the four-gluon vertex is given in Fig. 1.

You can verify that the three- and four-gluon vertices we introduced above are exactly those which arise from the Yang–Mills Lagrangian:

$$\mathcal{L}_{\rm YM} = -\frac{1}{4}\sum_a F^a_{\mu\nu}F^{a\mu\nu} \quad \text{with} \quad F^a_{\mu\nu}=\partial_{[\mu}A^a_{\nu]}-g\,f^{abc}A^b_{[\mu}A^c_{\nu]} . \tag{26}$$

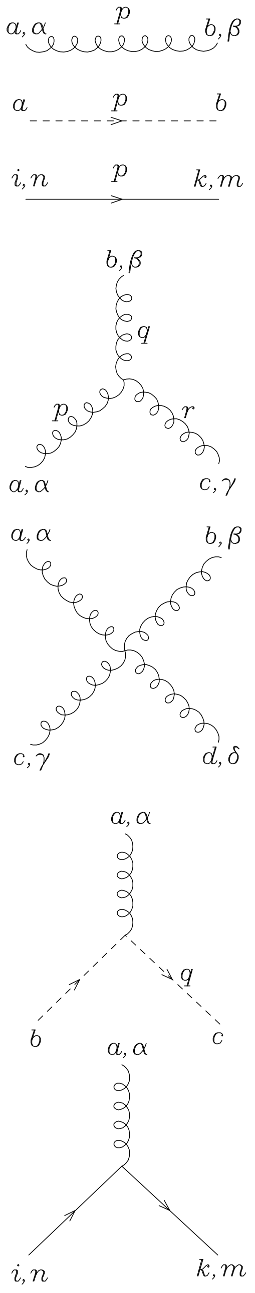
It can be shown that the three- and four-gluon vertices we generated are all is needed to guarantee gauge invariance even for processes more complicated than those studied in the previous simple examples. In other words, no extra five-or-more-gluon vertices have to be introduced to achieve the gauge invariance of higher-order amplitudes. At the tree level this is the consequence of dimensional analysis and of the locality of the couplings (no inverse powers of the momenta can appear in the Lagrangian). At the loop level, these conditions are supplemented by the renormalizability of the theory [3, 7].

Before one can start calculating cross-sections, a technical subtlety that arises in QCD when squaring the amplitudes and summing over the polarization of external states needs to be discussed. Let us again start from the QED example. Let us focus, for example, on the sum over polarizations of photon k_1:

$$\sum_{\epsilon_1}|M|^2 \;=\; \left(\sum_{\epsilon_1}\epsilon_1^{\mu}\epsilon_1^{\nu *}\right)M_\mu M_\nu^* . \tag{27}$$

The two independent physical polarizations of a photon with momentum $k=(k_0;0,0,k_0)$ are given by $\epsilon^\mu_{L,R}=(0;1,\pm i,0)/\sqrt{2}$. They satisfy the standard normalization properties:

$$\epsilon_L\cdot\epsilon_L^*=-1=\epsilon_R\cdot\epsilon_R^* \qquad \epsilon_L\cdot\epsilon_R^*=0$$



$$= \delta^{ab}\frac{-i\,g^{\alpha\beta}}{p^2+i\epsilon} \quad \text{(Feynman gauge)}$$

$$= \delta^{ab}\frac{i}{p^2+i\epsilon}$$

$$= \delta^{ik}\left.\frac{i}{\not p - m + i\epsilon}\right|_{mn}$$

$$= gf^{abc}\left[g^{\alpha\beta}(p-q)^{\gamma} + g^{\beta\gamma}(q-r)^{\alpha} + g^{\gamma\alpha}(r-p)^{\beta}\right]$$

$$\begin{aligned} = & -ig^2 f^{xac} f^{xbd}\left(g^{\alpha\beta}g^{\gamma\delta} - g^{\alpha\delta}g^{\beta\gamma}\right) \\ & -ig^2 f^{xad} f^{xbc}\left(g^{\alpha\beta}g^{\gamma\delta} - g^{\alpha\gamma}g^{\beta\delta}\right) \\ & -ig^2 f^{xab} f^{xcd}\left(g^{\alpha\gamma}g^{\beta\delta} - g^{\alpha\delta}g^{\beta\gamma}\right) \end{aligned}$$

$$= -gf^{abc}\,q^{\alpha}$$

$$= ig\,\lambda^{a}_{ki}\,\gamma^{\alpha}_{mn}$$

Fig. 1: Feynman rules for QCD. The solid lines represent the fermions, the curly lines the gluons, and the dotted lines the ghosts.

We can write the sum over physical polarizations in a convenient form by introducing the vector $\bar{k} = (k_0; 0, 0, -k_0)$:

$$\sum_{i=L,R} \epsilon_i^\mu \epsilon_i^{\nu *} \equiv \begin{pmatrix} 0 & & \vec{0} & \\ & 1 & 0 & 0 \\ \vec{0} & 0 & 1 & 0 \\ & 0 & 0 & 0 \end{pmatrix} = -g_{\mu\nu} + \frac{k_\mu \bar{k}_\nu + k_\nu \bar{k}_\mu}{k \cdot \bar{k}} \,. \tag{28}$$

We could have written the sum over physical polarizations using any other momentum ℓ_μ, provided $k \cdot \ell \neq 0$. This would be equivalent to a gauge transformation (prove it as an exercise). In QED the second term in Eq. (28) can be safely dropped, since $k_\mu M^\mu = 0$. As a cross-check, notice that $k_\mu M^\mu = 0$ implies $M_0 = M_3$, and therefore

$$\sum_{i=L,R} |\epsilon_i \cdot M|^2 = |M_1|^2 + |M_2|^2 = |M_1|^2 + |M_2|^2 + |M_3|^2 - |M_0|^2 \equiv -g^{\mu\nu}\, M_\mu M_\nu^* \,. \tag{29}$$

Therefore, the production of the longitudinal and time-like components of the photon cancel each other. This is true *regardless* of whether additional external photons are physical or not, since the gauge invariance $k_1 \cdot M = 0$ shown in Eq. (12) holds regardless of the choice for ϵ_2, as already remarked. In particular,

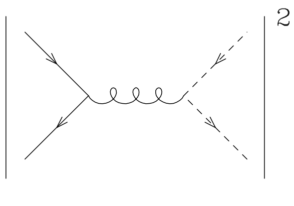
$$k_1^{\mu_1} k_2^{\mu_2} M_{\mu_1 \mu_2} = 0 \tag{30}$$

(for n photons, $k_1^{\mu_1} k_2^{\mu_2} \ldots k_n^{\mu_n}\, M_{\mu_1 \ldots \mu_n} = 0$) and the production of *any* number of unphysical photons vanishes. The situation in the case of gluon emission is different, since $k_1 \cdot M \propto \epsilon_2 \cdot k_2$, which vanishes only for a physical ϵ_2. This implies that the production of one physical gluon and one non-physical gluon is equal to 0, but the production of a pair of non-physical gluons is allowed! If $\epsilon_2 \cdot k_2 \neq 0$, then M_0 is not equal to M_3, and Eq. (28) is not equivalent to $\sum \epsilon_\mu \epsilon_\nu^* = -g_{\mu\nu}$.

Exercise: show that

$$\sum_{\text{non-physical}} |\epsilon_1^\mu \epsilon_2^\nu M_{\mu\nu}|^2 = \left| i\, g^2\, f^{abc} \lambda^c\, \frac{1}{2k_1 k_2} \bar{v}(\bar{q}) \not{k}_1\, u(q) \right|^2 . \tag{31}$$

In the case of non-Abelian theories, it is therefore important to restrict the sum over polarizations and (because of unitarity) the off-shell propagators to physical degrees of freedom with the choice of physical gauges. Alternatively, one has to undertake a study of the implications of gauge-fixing in non-physical gauges for the quantization of the theory (see Refs. [3,7]). The outcome of this analysis is the appearance of two colour-octet scalar degrees of freedom (called ghosts) whose rôle is to enforce unitarity in non-physical gauges. They will appear in internal closed loops, or will be pair-produced in final states. They only couple to gluons. Their Feynman rules are supplemented by the prescription that each closed loop should come with a -1 sign, as if they obeyed Fermi statistics. Being scalars, this prescription breaks the spin-statistics relation, and leads as a result to the possibility that production probabilities are negative. This is precisely what is required to cancel the contributions of non-transverse degrees of freedom appearing in non-physical gauges. Adding the ghosts contribution to $q\bar{q} \to gg$ decays (using the Feynman rules from Fig. 1) gives in fact



$$= -\left| i\, g^2\, f^{abc} \lambda^c\, \frac{1}{2k_1 k_2} \bar{v}(\bar{q}) \not{k}_1\, u(q) \right|^2 , \tag{32}$$

which exactly cancels the contribution of non-transverse gluons in the non-physical gauge $\sum\, \epsilon_\mu \epsilon_\nu^* = -g_{\mu\nu}$, given in in Eq. (31).

The detailed derivation of the need for and properties of ghosts (including their Feynman rules and the '-1' prescription for loops) can be found in the suggested textbooks. I will not derive these results here since we will not need them for our applications (we will use physical gauges or will consider processes not involving the three-gluon vertex). The full set of Feynman rules for the QCD Lagrangian is summarized in Fig. 1. Their application to the renormalization of vertices and couplings at one-loop is discussed in Appendix A.

2.2 Some useful results in colour algebra

The presence of colour factors in the Feynman rules makes it necessary to develop some technology to evaluate the colour coefficients which multiply our Feynman diagrams. To be specific, we shall assume the gauge group is SU(N). The fundamental relation of the algebra is

$$[\lambda^a, \lambda^b] = i f^{abd} \lambda^c \tag{33}$$

with f^{abc} totally antisymmetric. This relation implies that all λ matrices are traceless. For practical calculations, since we will always sum over initial, final, and intermediate state colours, we will never need the explicit values of f^{abc}. All of the results can be expressed in terms of group invariants (a.k.a. Casimirs), some of which we will now introduce. The first such invariant ($T_{\rm F}$) is chosen to fix the normalization of the matrices λ:

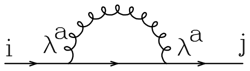
$$\mathrm{tr}(\lambda^a \lambda^b) = T_{\rm F} \delta_{ab} \tag{34}$$

where by convention $T_{\rm F} = 1/2$ for the fundamental representation. Should you change this convention, you would need to change the definition (i.e., the numerical value) of the coupling constant g, since $g\,\lambda^a$ appears in the Lagrangian and in the Feynman rules.

Exercise: Show that $\mathrm{tr}(\lambda^a \lambda^b)$ is indeed a group invariant. Hint: write the action on λ^a of a general group transformation with infinitesimal parameters ϵ^b as follows:

$$\delta \lambda^a = \sum_{b,c} \epsilon_b f^{abc} \lambda^c \, . \tag{35}$$

The definition of $T_{\rm F}$ allows us to evaluate the colour factor for an interesting diagram, i.e., the quark self-energy:



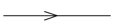
$$\sim \sum_a (\lambda^a \lambda^a)_{ij} \equiv C_{\rm F} \delta_{ij} \, . \tag{36}$$

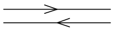
The value of $C_{\rm F}$ can be obtained by tracing the relation above:

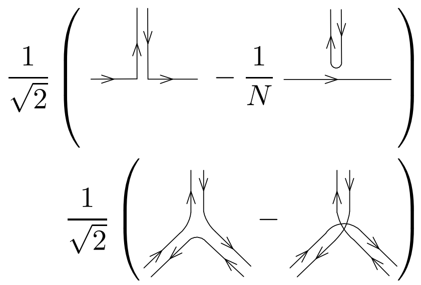
$$C_{\rm F} N \; = \; \mathrm{tr} \sum_a \lambda^a \lambda^a = \delta^{ab} \, T_{\rm F} \delta_{ab} = \frac{N^2-1}{2} \tag{37}$$

where we used the fact that $\delta^{ab}\delta_{ab} = N^2 - 1$, the number of matrices λ^a (and of gluons) for SU(N).

There are some useful graphical tricks (which I learned from P. Nason [9]) that can be used to evaluate complicated expressions. The starting point is the following representation for the quark and gluon propagators, and for the $q\bar{q}g$ and ggg interaction vertices:

 fermion (38)

 gluon (39)

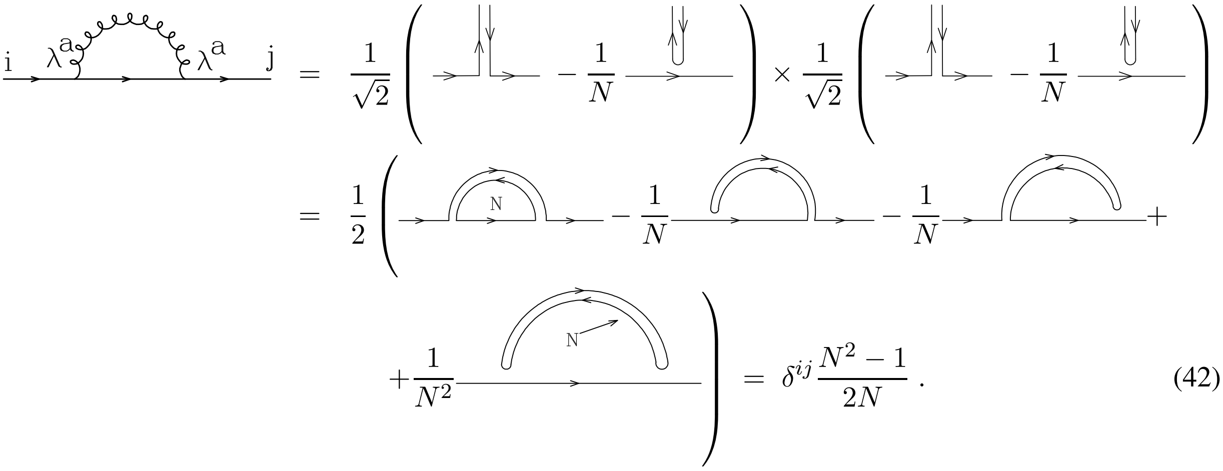


fermion–gluon vertex (t^a) (40)

3-gluon vertex (f^{abc}) . (41)

Contraction over colour indices is obtained by connecting the respective colour (or anticolour) lines. A closed loop of a colour line gives rise to a factor N, since the closed loop is equivalent to the trace of the unit matrix. So the above representation of the $q\bar{q}g$ vertex embodies the idea of ‘colour conservation’, whereby the colour–anticolour quantum numbers carried by the $q\bar{q}$ pair are transferred to the gluon. The piece proportional to $1/N$ in the $q\bar{q}g$ vertex appears only when the colour of the quark and of the antiquark are the same. It ensures that λ^a is traceless, as it should be. This can be easily checked as an exercise. The factor $1/\sqrt{2}$ is related to the chosen normalization of T_{F}.

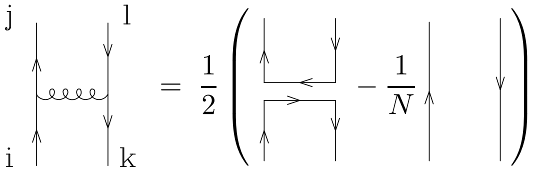
As a first example of applications, let us re-evaluate C_{F}:



$$= \delta^{ij}\frac{N^2-1}{2N} \ . \qquad (42)$$

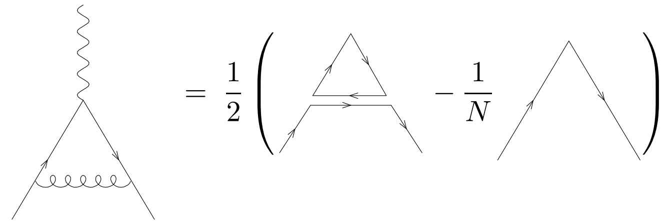
As an exercise, you can calculate the colour factor for $q\bar{q} \to q\bar{q}$ scattering, and show that

$$\sum_a (\lambda^a)_{ij}(\lambda^a)_{\ell k} =$$



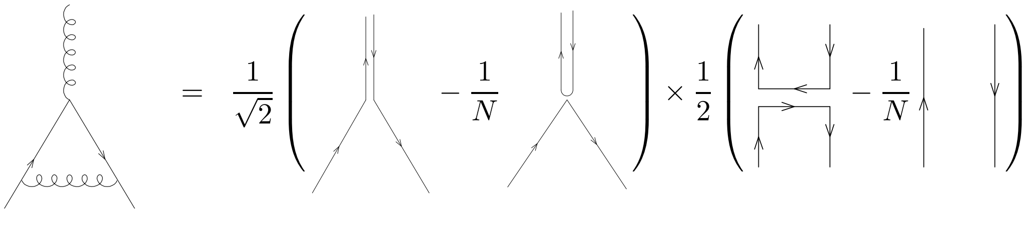
$$= \frac{1}{2}\left(\delta_{ik}\delta_{\ell j} - \frac{1}{N}\delta_{ij}\delta_{\ell k}\right) \ . \qquad (43)$$

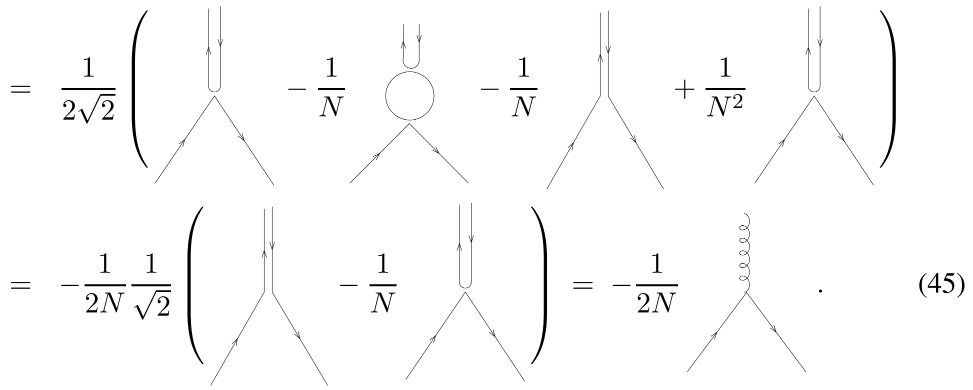
This result can be used to evaluate the one-loop colour factors for the interaction vertex with a photon:



$$= \frac{1}{2}\frac{N^2-1}{N}\delta_{ij} = C_{\mathrm{F}}\delta_{ij} \ . \qquad (44)$$

For the interaction with a gluon, we have instead





Note that in the case of the coupling to the photon, the $q\bar{q}$ pair is in a colour-singlet state and the gluon exchange effect has a positive sign (⇒ attraction). In the case of the coupling to the gluon, the $q\bar{q}$ pair is in a colour-octet state and the gluon-exchange correction has a negative sign relative to the Born interaction. The force between a $q\bar{q}$ pair is therefore *attractive* if the pair is in a colour-singlet state, and *repulsive* if it is in a colour-octet state! This gives a qualitative argument for why no colour-octet $q\bar{q}$ bound state exists.

The remaining important relation that one needs is the following:

$$\sum_{a,b} f^{abc} f^{abd} = C_{\mathrm{A}} \delta^{cd} \text{ with } C_{\mathrm{A}} = N \, . \tag{46}$$

You can easily prove it by using the graphical representation given in Eq. (41), or by using Eq. (43) and $f^{abc} = -2i \,\mathrm{tr}([\lambda^a, \lambda^b]\, \lambda^c)$.

3 QCD and the proton structure at large Q^2

The understanding of the structure of the proton at short distances is one of the key ingredients to be able to predict cross-section for processes involving hadrons in the initial state. All processes in hadronic collisions, even those intrinsically of electroweak nature such as the production of W/Z bosons or photons, are in fact induced by the quarks and gluons contained inside the hadron. In this lecture I will introduce some important concepts, such as the notion of partonic densities of the proton, and of parton evolution. These are the essential tools used by theorists to predict production rates for hadronic reactions.

We shall limit ourselves to processes where a proton–(anti)proton pair collides at high centre-of-mass energy ($\sqrt{S}$, typically larger than several hundred GeV) and undergoes a strongly inelastic interaction, with momentum transfers between the participants in excess of several GeV. The outcome of this hard interaction could be the simple scattering at large angle of some of the hadron's elementary constituents, their annihilation into new massive resonances, or a combination of the two. In all cases the final state consists of a large multiplicity of particles, associated with the evolution of the fragments of the initial hadrons, as well as of the new states produced. As discussed below, the fundamental physical concept that makes this programme possible is 'factorization', the ability to isolate separate independent phases of the overall collision. These phases are dominated by different dynamics, and the most appropriate techniques can be applied to describe each of them separately. In particular, factorization allows one to decouple the complexity of the proton structure and of the final-state hadron formation from the elementary nature of the perturbative hard interaction among parton constituents.

Figure 2 illustrates how this works. As the left proton travels freely before coming into contact with the hadron coming in from the right, its constituent quarks are held together by the constant exchange of virtual gluons (e.g., gluons a and b in the picture). These gluons are mostly soft, because any hard exchange would cause the constituent quarks to fly apart, and a second hard exchange would be necessary to re-establish the balance of momentum and keep the proton together. Gluons of high virtuality (gluon c in the picture) prefer therefore to be reabsorbed by the same quark, within a time inversely

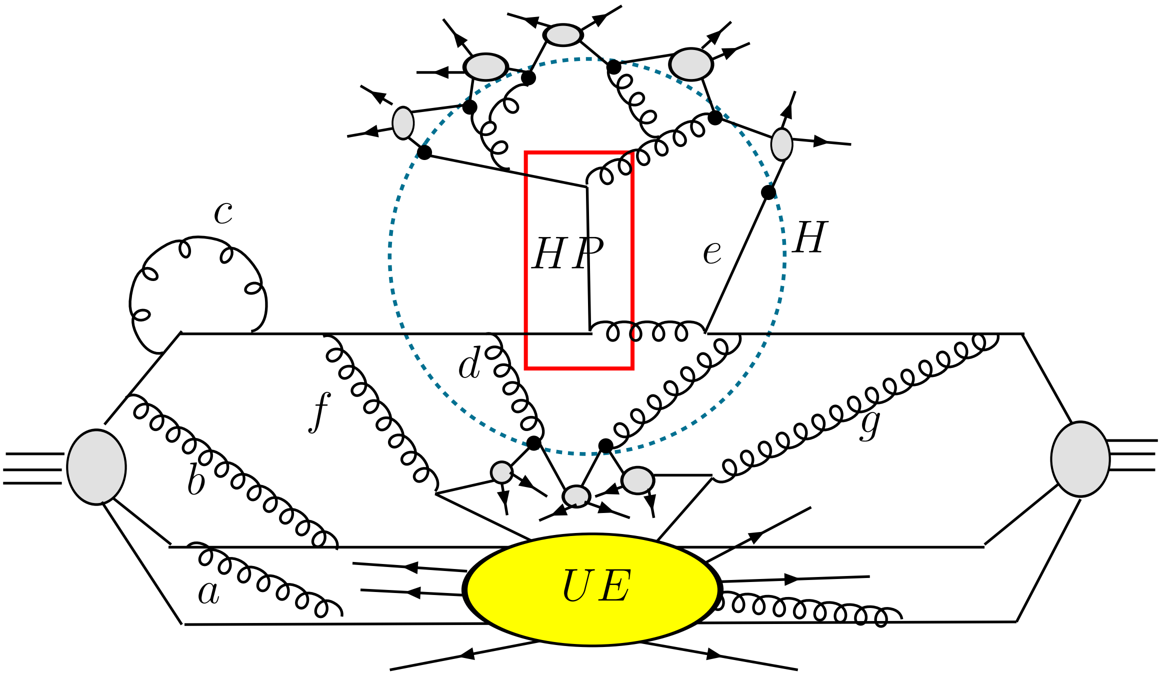


Fig. 2: General structure of a hard proton–proton collision

proportional to their virtuality, as prescribed by the uncertainty principle. The state of the quark is, however, left unchanged by this process. Altogether this suggests that the global state of the proton, although defined by a complex set of gluon exchanges between quarks, is nevertheless determined by interactions which have a time scale of the order of $1/m_p$. When seen in the laboratory frame where the proton is moving with energy $\sqrt{S}/2$, this time is furthermore Lorentz dilated by a factor $\gamma = \sqrt{S}/2m_p$. If we disturb a quark with a probe of virtuality $Q \gg m_p$, the time frame for this interaction is so short ($1/Q$) that the interactions of the quark with the rest of the proton can be neglected. The struck quark cannot negotiate with its partners a coherent response to the external perturbation: it simply does not have the time to communicate to them that it is being kicked away. On this time scale, only gluons with energy of the order of Q can be emitted, something which, to happen coherently over the whole proton, is suppressed by powers of m_p/Q (this suppression characterizes the 'elastic form factor' of the proton). In this figure, the hard process is represented by the rectangle labelled HP. In this example a head-on collision with a gluon from the opposite hadron, leads to a $qg \to qg$ scattering with a momentum exchange of the order of Q. This and other possible processes can be calculated from first principles in perturbative QCD.

When the constituent is suddenly deflected, the partons that it had recently radiated cannot be reabsorbed (as happened to gluon c earlier) because the constituent is no longer there waiting for the partons to come back. This is the case, for example, of the gluon d emitted by the quark, and of the quark e from the opposite hadron; the emitted gluon got engaged in the hard interaction. The number of 'liberated' partons will depend on the hard scale Q: the larger Q, the more sudden the deflection of the struck parton, and the fewer the partons that can reconnect before its departure (typically only partons with virtuality larger than Q).

After the hard process, the partons liberated during the evolution prior to the collision and the partons created by the hard collision will themselves emit radiation. The radiation process, governed by perturbative QCD, continues until a low virtuality scale is reached (the boundary region labelled with a dotted line, H, in our figure). To describe this perturbative evolution phase, proper care has to be taken to incorporate quantum coherence effects, which in principle connect the probabilities of radiation off different partons in the event. ăOnce the low virtuality scale is reached the memory of the hard-process phase has been lost, once again as a result of different time scales in the problem, and the final phase of hadronization takes over. Because of the decoupling from the hard-process phase, the hadronization is assumed to be independent of the initial hard process, and its parametrization, tuned to the observables of some reference process, can then be used in other hard interactions (universality of hadronization). Nearby partons merge into colour-singlet clusters (the grey blobs in Fig. 2), which are then decayed phenomenologically into physical hadrons. To complete the picture, we need to understand the evolution of the fragments of the initial hadrons. As shown in the figure, this evolution cannot be entirely independent of what happens in the hard event, because at least colour quantum numbers must be exchanged to guarantee the overall neutrality and conservation of baryon number. In our example, the gluons f and g, emitted early on in the perturbative evolution of the initial state, split into $q\bar{q}$ pairs which are shared between the hadron fragments (whose overall interaction is represented by the oval labelled UE, for Underlying Event) and the clusters resulting from the evolution of the initial state.

The above ideas are embodied in the following factorization formula, which represents the starting point of any theoretical analysis of cross-sections and observables in hadronic collisions:

$$\frac{d\sigma}{dX} = \sum_{j,k} \int_{\hat{X}} f_j(x_1, Q_i) f_k(x_2, Q_i) \, \frac{d\hat{\sigma}_{jk}(Q_i, Q_f)}{d\hat{X}} \, F(\hat{X} \to X;\, Q_i, Q_f) \, , \tag{47}$$

where

- X is some hadronic observable (e.g., the transverse momentum of a pion);
- the sum over j and k extends over the parton types inside the colliding hadrons;
- the function $f_j(x, Q)$ (known as parton distribution function, PDF) represents the number density of parton type j with momentum fraction x in a proton probed at a scale Q_i;
- $\hat{X}$ is a parton-level observable (e.g., the transverse momentum of a parton from the hard scattering);
- $\hat{\sigma}_{jk}$ is the parton-level cross-section, differential in the observable $\hat{X}$;
- $F(\hat{X} \to X;\, Q_i, Q_f)$ is a transition function, weighting the probability that the partonic state defining $\hat{X}$ gives rise to the hadronic observable X;
- the scales Q_i and Q_f correspond to the scales at which we separate the perturbative, hard process from the initial- and final-state evolutions, respectively.

In the rest of this section I shall cover the above ideas in some more detail. While I will not provide you with a rigorous proof of the legitimacy of this approach, I will try to justify it qualitatively to make it sound at least plausible. In Appendix B I will collect some more explicit derivations and results.

3.1 The parton densities and their evolution

As mentioned above, the binding forces responsible for the quark confinement are due to the exchange of rather soft gluons. If a quark were to exchange a hard virtual gluon with another quark, in fact, the recoil would tend to break the proton apart. It is easy to verify that the exchange of gluons with virtuality larger than Q is then proportional to some large power of m_p/Q, m_p being the proton mass. Since the gluon coupling constant gets smaller at large Q, exchange of hard gluons is significantly suppressed [1]. Consider in fact the picture Fig. 3. The exchange of two gluons is required to ensure that the momentum exchanged after the first gluon emission is returned to the quark, and the proton maintains its structure. The contributions of hard gluons to this process can be approximated by integrating the loop over large momenta:

$$\int_Q \frac{d^4q}{q^6} \sim \frac{1}{Q^2} \, . \tag{48}$$

[1] The fact that the coupling decreases at large Q plays a fundamental role in this argument. Were this not true, the parton picture could not be used!

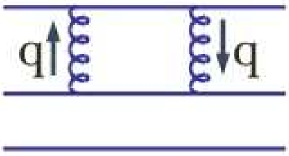


Fig. 3: Gluon exchange inside the proton

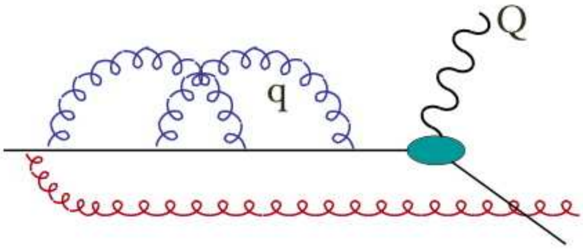


Fig. 4: Gluon emission at different scales during the approach to a hard collision

At large Q this contribution is suppressed by powers of $(m_p/Q)^2$, where the proton mass m_p is included as being the only dimensionful quantity available. The interactions keeping the proton together are therefore dominated by soft exchanges, with virtuality Q of the order of m_p. The typical time scale of these exchanges is of the order of $1/m_p$ (in the laboratory system, where the proton travels with energy E, this time is Lorentz dilated to $\tau \sim \gamma/m_p = E/m_p^2$). If we probe the proton with an off-shell photon, the interaction takes place during the limited lifetime of the virtual photon, given by the inverse of its virtuality as a result of the Heisenberg principle. Assuming the virtuality $Q \gg m_p$, once the photon gets 'inside' the proton and meets a quark, the struck quark has no time to negotiate a coherent response with the other quarks, because the time scale for it to 'talk' to its pals is too long compared with the duration of the interaction with the photon itself. As a result, the struck quark has no option but to interact with the photon as if it were a free particle.

Let us look in more detail at what happens during such process. In Fig. 4 we see a proton as it approaches a hard collision with a photon of virtuality Q. Gluons emitted at a scale $q > Q$ have the time to be reabsorbed, since their lifetime is very short. Their contribution to the process can be calculated in perturbative QCD, since the scale is large. Since after being reabsorbed the state of the quark remains the same, their only effect is an overall renormalization of the wave function, and they do not affect the quark density. A gluon emitted at a scale $q < Q$, however, has a lifetime longer than the time it takes for the quark to interact with the photon, and by the time it tries to reconnect to its parent quark, the quark has been kicked away by the photon, and is no longer there. Since the gluon has taken away some of the quark momentum, the momentum fraction x of the quark as it enters the interaction with the photon is different from the momentum it had before, and therefore its density $f(x)$ is affected. Furthermore, when the scale q is of the order of 1 GeV the state of the quark is not calculable in perturbative QCD. This state depends on the internal wave function of the proton, which perturbative QCD cannot easily predict. We can however say that the wave function of the proton, and therefore the state of the 'free' quark, are determined by the dynamics of the soft-gluon exchanges inside the proton itself. Since the time scale of this dynamics is long relative to the time scale of the photon–quark interaction, we can safely argue that the photon sees to good approximation a static snapshot of the proton's inner guts. In other words, the state of the quark had been prepared long before the photon arrived. This also suggests that the state of the quark will not depend on the precise nature of the external probe, provided the time scale of the hard interaction is very short compared to the time it would take for the quark to readjust itself. As a result, if we could perform some measurement of the quark state using, say, a virtual-photon probe, we could then use this knowledge on the state of the quark to perform predictions for the interaction of the proton

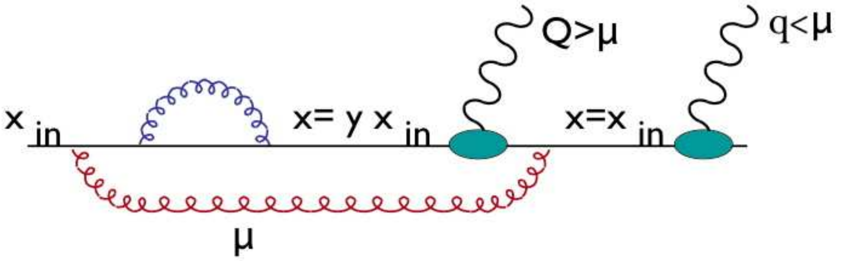


Fig. 5: Scale dependence of the gluon emission during a hard collision

with any other probe (e.g., a virtual W or even a gluon from an opposite beam of hadrons). This is the essence of the universality of the parton distributions.

The above picture leads to an important observation. It appears in fact that which gluons are reabsorbed and which ones are not depends on the scale Q of the hard probe. As a result, the parton density $f(x)$ appears to depend on Q. This is illustrated in Fig. 5. The gluon emitted at a scale μ has a lifetime short enough to be reabsorbed before a collision with a photon of virtuality $Q < \mu$, but too long for a photon of virtuality $Q > \mu$. When going from μ to Q, therefore, the partonic density $f(x)$ changes. We can easily describe this variation as follows:

$$f(x,Q) \; = \; f(x,\mu) \, + \, \int_x^1 dx_{\rm in} \, f(x_{\rm in},\mu) \, \int_\mu^Q dq^2 \, \int_0^1 dy \, \mathcal{P}(y,q^2) \, \delta(x - y x_{\rm in}) \, . \qquad (49)$$

Here we obtain the density at the scale Q by adding to $f(x)$ at the scale μ (which we label $f(x,\mu)$) all the quarks with momentum $x_{\rm in} > x$, which retain a momentum fraction $x = y/x_{\rm in}$ by emitting a gluon. The function $P(y,Q^2)$ describes the 'probability' that the quark emits a gluon at a scale Q, keeping a fraction y of its momentum. This function does not depend on the details of the hard process, it simply describes the radiation of a quark subject to an interaction with virtuality Q. Since $f(x,Q)$ does not depend upon μ (μ is just used as a reference scale to construct our argument), the total derivative of the right-hand side with respect to μ should vanish, leading to the following equation:

$$\frac{df(x,Q)}{d\mu^2} = 0 \quad \Rightarrow \frac{df(x,\mu)}{d\mu^2} \; = \; \int_x^1 \frac{dy}{y} \, f(y,\mu) \, \mathcal{P}(x/y,\mu^2) \, . \qquad (50)$$

One can prove (see Appendix B) that

$$\mathcal{P}(x,Q^2) \; = \; \frac{\alpha_s}{2\pi} \frac{1}{Q^2} \, P(x) \; , \qquad (51)$$

from which the Altarelli–Parisi equation follows:

$$\frac{df(x,\mu)}{d\log\mu^2} \; = \; \frac{\alpha_s}{2\pi} \int_x^1 \frac{dy}{y} \, f(y,\mu) \, P_{qq}(x/y) \, . \qquad (52)$$

The so-called splitting function $P_{qq}(x)$ can be calculated in perturbative QCD, and is given in Appendix B. The subscript qq is a convention indicating that x refers to the momentum fraction retained by a quark after emission of a gluon.

More in general, one should consider additional processes. For example, one should include cases in which the quark interacting with the photon comes from the splitting of a gluon. This is shown in Fig. 6: the left diagram is the one we considered above; the right diagram corresponds to processes where an emitted gluon has the time to split into a $q\bar{q}$ pair, and it is one of these quarks which interacts with the photon. The overall evolution equation, including the effect of gluon splitting, is given by

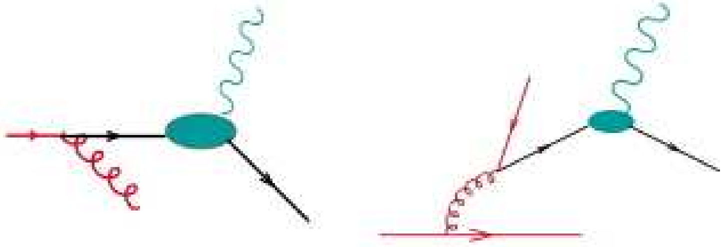


Fig. 6: The processes leading to the evolution of the quark density

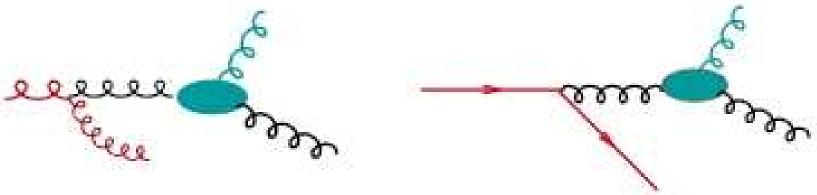


Fig. 7: The processes leading to the evolution of the gluon density

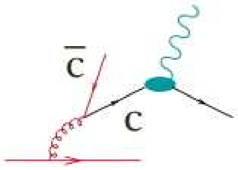


Fig. 8: Gluon evolution leading to a charm-quark content of the proton

$$\frac{dq(x,Q)}{dt} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dy}{y} \left[q(y,Q)\, P_{qq}(\frac{x}{y}) \;+\; g(y,Q)\, P_{qg}(\frac{x}{y}) \right] . \tag{53}$$

For external probes which couple to gluons (namely an external gluon, coming, for example, from an incoming proton), we have a similar evolution of the gluon density (see Fig. 7):

$$\frac{dg(x,Q)}{dt} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dy}{y} \left[g(y,Q)\, P_{gg}(\frac{x}{y}) \;+\; \sum_{q,\bar{q}} q(y,Q)\, P_{gq}(\frac{x}{y}) \right] . \tag{54}$$

3.2 Example: the charm content of the proton

If the virtuality of the external probe is large enough, the time scale of the hard interaction is so short that gluon fluctuations into virtual heavy-quark states can be intercepted, and the virtual heavy quarks (charm quarks in our example) can be brought on-shell via the interaction with the photon (see Fig. 8). To the external photon, it will therefore appear as if the proton contained some charm. Its density can be calculated using the Altarelli–Parisi equation, assuming that the heavy-quark density itself is 0 at $Q \sim m_c$, and builds up according to the evolution equation

$$\frac{dc(x,Q)}{dt} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dy}{y}\, g(y,Q)\, P_{qg}(\frac{x}{y}) . \tag{55}$$

Assuming a gluon density behaving like $g(x,Q) \sim A/x$, which is a first approximation to a bremsstrahlung spectrum, we can easily calculate

$$\frac{dc(x,Q)}{dt} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dy}{y}\, g(x/y,Q)\, P_{qg}(y) = \frac{\alpha_s}{2\pi} \int_x^1 dy\, \frac{A}{x}\, \frac{1}{2}\, [y^2 + (1-y)^2]$$

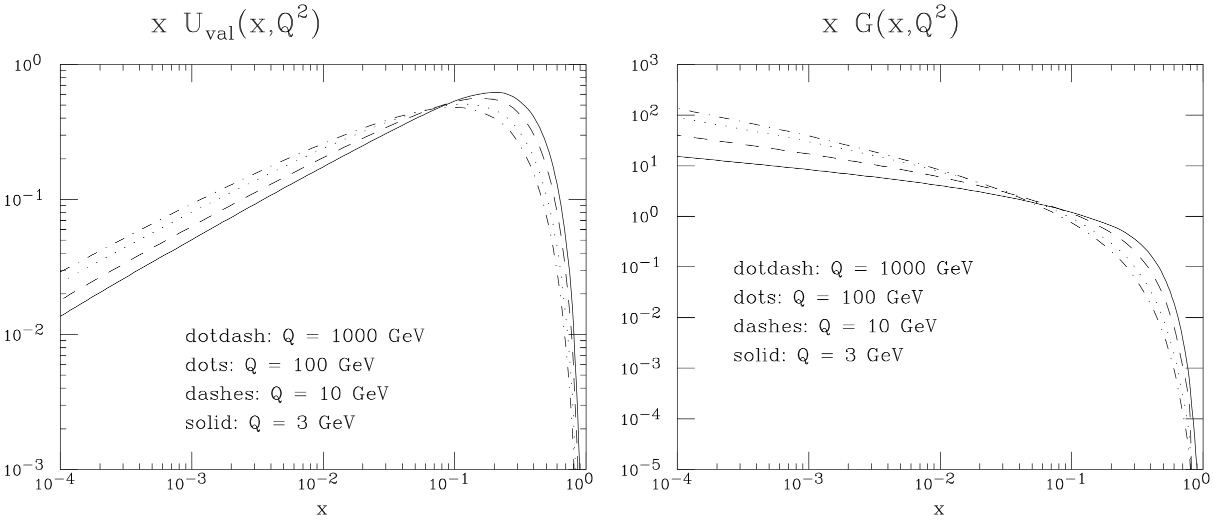


Fig. 9: Left: Valence up-quark momentum-density distribution, for different scales Q. Right: gluon momentum-density distribution.

$$= \frac{\alpha_s}{6\pi}\frac{A}{x}c(x,Q) \sim \frac{\alpha_s}{6\pi}\log(\frac{Q^2}{m_c^2})\, g(x,Q)\,. \tag{56}$$

The charm density is therefore proportional to the gluon density, up to an overall factor proportional to α_s. When Q becomes very large, the effect of the quark mass becomes subleading, and we expect all sea quarks to reach asymptotically the same density!

3.3 Examples of parton density evolution

Figure 9 (left) describes the up-quark valence momentum density at different scales Q. Note the softening at large scales, and the clear logQ^2 evolution. As Q^2 grows, the valence quarks emit more and more radiation, since their deceleration is larger. They therefore lose more momentum to the emitted gluons, and their spectrum becomes softer. The most likely momentum fraction carried by a valence up-quark in the proton goes from $x \sim 20\%$ at $Q = 3$ GeV, to $x \lesssim 10\%$ at $Q = 1000$ GeV. Notice finally that the density vanishes at small x.

Figure 9 (right) shows the gluon momentum density at different scales Q. Their density grows at small x, with an approximate $g(x) \sim 1/x^{1+\delta}$ behaviour, and $\delta > 0$ slowly increasing at large Q^2. This low-x growth is due to the $1/x$ emission probability for the radiation of gluons, which was discussed in the previous lecture and which is represented by the $1/x$ factors in the $P_{gq}(x)$ and $P_{gg}(x)$ splitting functions. As Q^2 grows we find an increasing number of gluons at small x, as a result of the increased radiation off quarks, as well as off the harder gluons.

Figure 10 (left) shows the evolution of the up-quark *sea* momentum density. Shape and evolution match those of the gluon density, a consequence of the fact that sea quarks come from the splitting of gluons. Since the gluon-splitting probability is proportional to α_s, the approximate ratio sea/gluon ~ 0.1 which can be obtained by comparing Figs. 9 (right) and 10 (left) is perfectly justified.

Finally, the momentum densities for gluons, up-sea, charm and up-valence distributions are shown in Fig. 10 (right) for $Q = 1000$ GeV. Note here that u_{sea} and charm are approximately the same at very large Q and small x, as anticipated in the previous subsection. The proton momentum is mostly carried by valence quarks and by gluons. The contribution of sea quarks is negligible.

Parton densities are extracted from experimental data. Their determination is therefore subject to the statistical and systematic uncertainties of the experiments and of the theoretical analysis (e.g., the treatment of non-perturbative effects, the impact of missing higher-order perturbative corrections).

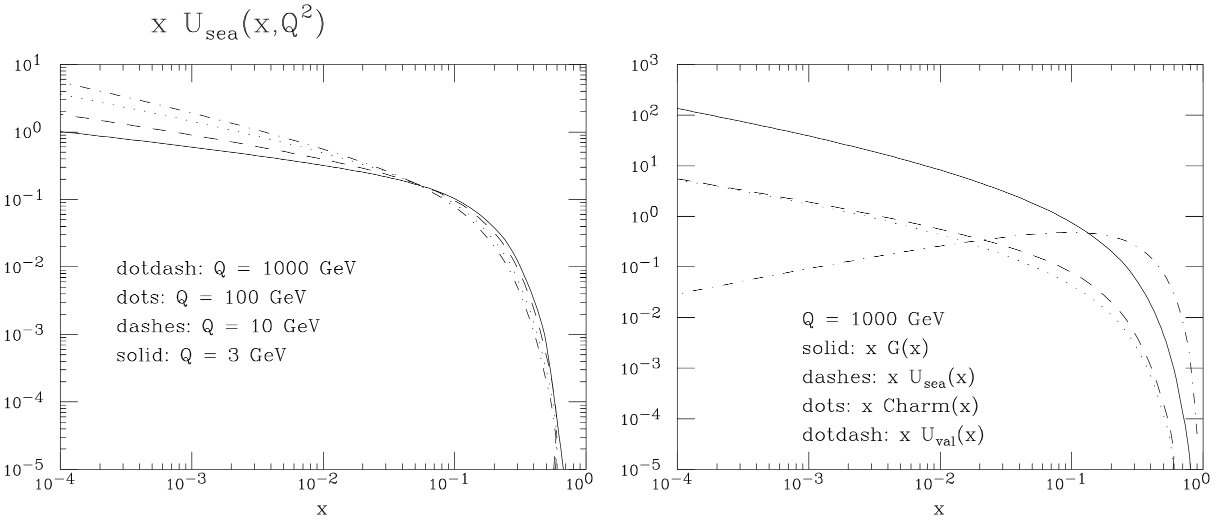


Fig. 10: Left: Sea up-quark momentum-density distribution, for different scales Q. Right: Momentum-density distribution for several parton species, at $Q = 1000$ GeV.

Techniques have been introduced recently to take into account these uncertainties, and to evaluate their impact on concrete observables. A summary of such an analysis is given in Figs. 11 (for the Tevatron) and 12 (for the LHC). What is plotted is the uncertainty bands for partonic luminosities corresponding to various initial-state channels, such as gg, qg or $q\bar{q}$. The partonic flux is given as a function of $\hat{s}$, the partonic centre-of-mass invariant mass. Obvious features include the growth of uncertainty of the gg density at large mass, corresponding to the lack of data covering the large-x region of the gluon density. As a result of this, notice for example that the uncertainty in the $gg \to t\bar{t}$ production rate at the LHC is smaller than at the Tevatron, since the relative range of mass (just above $2m_t \sim 350$ GeV) corresponds at the LHC to gluon densities in better explored regions of x.

4 The evolution of quarks and gluons

We discussed in the previous section the initial-state evolution of quarks and gluons as the proton approaches the hard collision. We study here how quarks and gluons evolve, and finally transform into hadrons, neutralizing their colours. We start by considering the simplest case, e^+e^- collisions, which provide the cleanest environment in which to study applications of QCD at high energy. This is the place where theoretical calculations have today reached their best accuracy, and where experimental data are the most precise, especially thanks to the huge statistics accumulated by LEP, LEP2 and SLC. The key process is the annihilation of the e^+e^- pair into a virtual photon or Z^0 boson, which will subsequently decay to a $q\bar{q}$ pair. e^+e^- collisions have therefore the big advantage of providing an almost point-like source of quark pairs, so that, contrary to the case of interactions involving hadrons in the initial state, we at least know very precisely the state of the quarks at the beginning of the interaction process.

Nevertheless, it is by no means obvious that this information is sufficient to predict the properties of the hadronic final state. We know that this final state is clearly not simply a $q\bar{q}$ pair, but some high-multiplicity set of hadrons. For example, the average multiplicity of charged hadrons in the decay of a Z^0 is approximately 20! It is therefore not obvious that a calculation done using the simple picture $e^+e^- \to q\bar{q}$ will have anything to do with reality. For example, one may wonder why we do not need to calculate $\sigma(e^+e^- \to q\bar{q}g \ldots g \ldots)$ for all possible gluon multiplicities to get an accurate estimate of $\sigma(e^+e^- \to \text{hadrons})$. And since in any case the final state is not made of q's and g's, but of π's, K's, ρ's, etc., why would $\sigma(e^+e^- \to q\bar{q}g \ldots g)$ be enough?

The solution to this puzzle lies both in time and energy scales, and in the dynamics of QCD. When

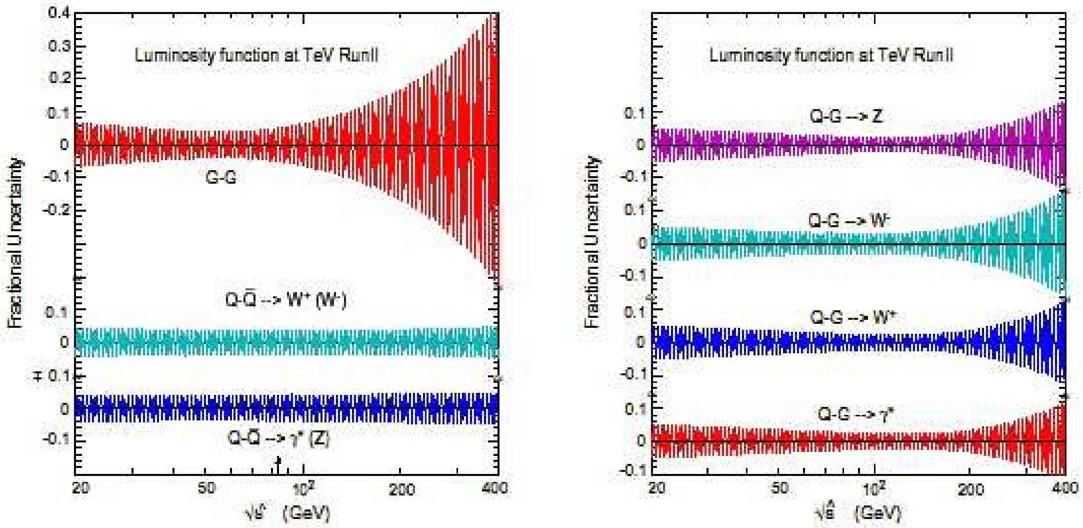


Fig. 11: Uncertainty in the parton luminosity functions at the Tevatron

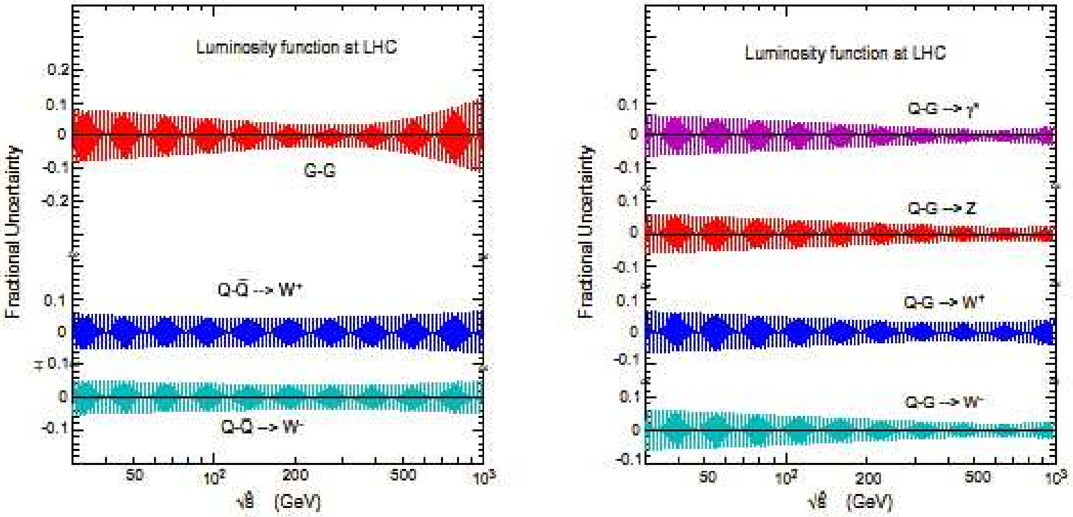


Fig. 12: Uncertainty in the parton luminosity functions at the LHC

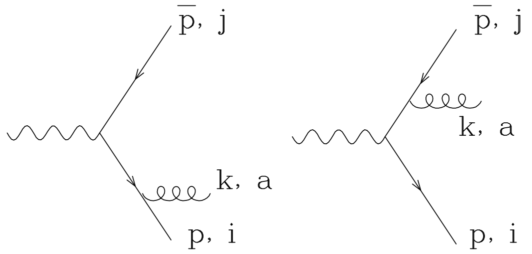
the $q\bar{q}$ pair is produced, the force binding q and $\bar{q}$ is proportional to $\alpha_s(s)$ ($\sqrt{s}$ being the e^+e^- centre-of-mass energy). Therefore it is weak, and q and $\bar{q}$ behave to good approximation like free particles. The radiation emitted in the first instants after the pair creation is also perturbative, and it will stay so until a time after creation of the order of $(1\ \mathrm{GeV})^{-1}$, when radiation with wavelengths $\gtrsim\ (1\ \mathrm{GeV})^{-1}$ starts being emitted. At this scale the coupling constant is large, and non-perturbative phenomena and hadronization start playing a role. However, as we will show, colour emission during the perturbative evolution organizes itself in such a way as to form colour-neutral, low-mass parton clusters highly localized in phase-space. As a result, the complete colour-neutralization (i.e., the hadronization) does not involve long-range interactions between partons far away in phase-space. This is very important, because the forces acting among coloured objects at this time scale would be huge. If the perturbative evolution were to separate far apart colour-singlet $q\bar{q}$ pairs, the final-state interactions taking place during the hadronization phase would totally upset the structure of the final state. As an additional result of this ‘pre-confining’ evolution, memory of where the local colour-neutral clusters came from is totally lost. So we expect the properties of hadronization to be universal: a model that describes hadronization at a given energy will work equally well at some other energy. Furthermore, so much time has passed since the original $q\bar{q}$ creation, that the hadronization phase cannot significantly affect the total hadron production rate. Perturbative corrections due to the emission of the first hard partons should be calculable in perturbation theory (PT), providing a finite, meaningful cross-section.

The nature of non-perturbative corrections to this picture can be explored. One can prove for example that the leading correction to the total rate $R_{e^+e^-}$ is of order F/s^2, where $F \propto \langle 0|\alpha_s F^a_{\mu\nu} F^{\mu\nu a}|0\rangle$ is the so-called gluon condensate. Since $F \sim \mathcal{O}(1\ \mathrm{GeV}^4)$, these NP corrections are usually very small. For example, they are of $\mathcal{O}(10^{-8})$ at the Z^0 peak! Corrections scaling like Λ^2/s or $\Lambda/\sqrt{s}$ can nevertheless appear in other less inclusive quantities, such as event shapes or fragmentation functions.

We now come back to the perturbative evolution, and will devote the first part of this lecture to justifying the picture given above. In Appendix C we shall discuss some applications, including jet cross-sections and shape variables.

4.1 Soft gluon emission

Emission of soft gluons plays a fundamental rôle in the evolution of the final state [6, 15]. Soft gluons are emitted with large probability, since the emission spectrum behaves like dE/E, typical of bremsstrahlung as familiar in QED. They provide the seed for the bulk of the final-state multiplicity of hadrons. The study of soft-gluon emission is simplified by the simplicity of their couplings. Being soft (i.e., long wavelength) they are insensitive to the details of the very-short-distance dynamics: they cannot distinguish features of the interactions which take place on time scales shorter than their wavelength. They are also insensitive to the spin of the partons: the only feature they are sensitive to is the colour charge. To prove this let us consider soft-gluon emission in the $q\bar{q}$ decay of an off-shell photon:



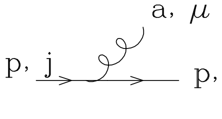
(57)

$$\begin{aligned} A_{\text{soft}} &= \bar{u}(p)\epsilon(k)(ig)\,\frac{-i}{\not{p}+\not{k}}\,\Gamma^\mu\, v(\bar{p})\ \lambda^a_{ij} \;+\; \bar{u}(p)\,\Gamma^\mu\,\frac{i}{\bar{\not{p}}+\not{k}}\,(ig)\epsilon(k)\, v(\bar{p})\ \lambda^a_{ij} \\ &= \left[\frac{g}{2p\cdot k}\,\bar{u}(p)\epsilon(k)\,(\not{p}+\not{k})\Gamma^\mu\, v(\bar{p}) \;-\; \frac{g}{2\bar{p}\cdot k}\,\bar{u}(p)\,\Gamma^\mu\,(\bar{\not{p}}+\not{k})\epsilon(k)\, v(\bar{p})\right]\lambda^a_{ij}\,. \end{aligned}$$

I used the generic symbol Γ_μ to describe the interaction vertex with the photon to stress the fact that the following manipulations are independent of the specific form of Γ_μ. In particular, Γ_μ can represent an arbitrarily complicated vertex form factor. Neglecting the factors of $\not{k}$ in the numerators (since $k \ll p, \bar{p}$, by definition of soft) and using the Dirac equations, we get

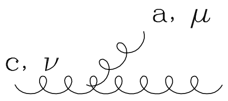
$$A_{\text{soft}} = g\lambda^a_{ij}\,\left(\frac{p\cdot\epsilon}{p\cdot k} - \frac{\bar{p}\epsilon}{\bar{p}\cdot k}\right)\, A_{\text{Born}}\,. \tag{58}$$

We then conclude that soft-gluon emission factorizes into the product of an emission factor, times the Born-level amplitude. From this exercise, one can extract general Feynman rules for soft-gluon emission:



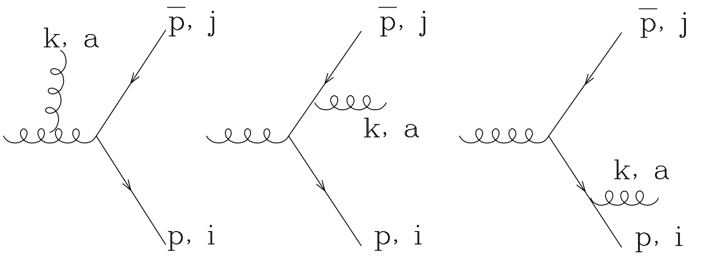
$$= g\,\lambda^a_{ij}\,2p^\mu\,. \tag{59}$$

Exercise: Derive the $g \to gg$ soft-emission rules:



$$= igf^{abc}\, 2p^{\mu}\, g^{\nu\rho}\ . \tag{60}$$

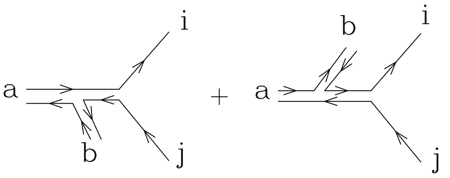
Example: Consider the ‘decay’ of a virtual gluon into a quark pair. One more diagram should be added to those considered in the case of the electroweak decay. The fact that the quark pair is no longer in a colour-singlet state makes things a bit more interesting:



(61)

$$\begin{aligned} &\overset{k\to 0}{=} \left[i\, g\, f^{abc}\, \lambda^c_{ij} \left(\frac{Q\epsilon}{Qk}\right) + g\, (\lambda^b\, \lambda^a)_{ij} \left(\frac{p\epsilon}{pk}\right) - g\, (\lambda^a \lambda^b)_{ij} \left(\frac{\bar p\epsilon}{pk}\right) \right] A_{\rm Born} \\ &= \quad g\, (\lambda^a\, \lambda^b)_{ij} \left[\frac{Q\epsilon}{Qk} - \frac{\bar p\epsilon}{pk}\right] + g\, (\lambda^b\, \lambda^a)_{ij} \left[\frac{p\epsilon}{pk} - \frac{Q\epsilon}{Qk}\right]\ . \end{aligned} \tag{62}$$

The two factors correspond to the two possible ways colour can flow in this process:



(63)

In the first case, the antiquark (colour label j) is colour connected to the soft gluon (colour label b) and the quark (colour label i) is connected to the decaying gluon (colour label a). In the second case, the order is reversed. The two emission factors correspond to the emission of the soft gluon from the antiquark, and from the quark line, respectively. When squaring the total amplitude, and summing over initial- and final-state colours, the interference between the two pieces is suppressed by $1/N^2$ relative to the individual squares:

$$\sum_{a,b,i,j} |(\lambda^a \lambda^b)_{ij}|^2 = \sum_{a,b} {\rm tr}\, \left(\lambda^a \lambda^b \lambda^b \lambda^a\right) = \frac{N^2-1}{2}\, C_{\rm F} \; = \; \mathcal{O}(N^3) \tag{64}$$

$$\sum_{a,b,i,j} (\lambda^a \lambda^b)_{ij} [(\lambda^b \lambda^a)_{ij}]^* = \sum_{a,b} {\rm tr}(\lambda^a \lambda^b \lambda^a \lambda^b) = \frac{N^2-1}{2}\, \underbrace{(C_{\rm F} - \frac{C_{\rm A}}{2})}_{-\frac{1}{2N}} \; = \; \mathcal{O}(N)\ . \tag{65}$$

As a result, the emission of a soft gluon can be described, to the leading order in $1/N^2$, as the incoherent sum of the emission from the two colour currents.

4.2 Angular ordering for soft-gluon emission

The results presented above have important consequences for the perturbative evolution of the quarks. A key property of the soft-gluon emission is the so-called *angular ordering*. This phenomenon consists in the continuous reduction of the opening angle at which successive soft gluons are emitted by the evolving quark. As a result, this radiation is confined within smaller and smaller cones around the quark direction, and the final state will look like a collimated jet of partons. In addition, the structure of the colour flow during the jet evolution forces the $q\bar{q}$ pairs which are in a colour-singlet state to be close in phase-space, thereby achieving the pre-confinement of colour-singlet clusters alluded to at the beginning of this section.

Let us start by proving the property of colour ordering. Consider the $q\bar{q}$ pair produced by the decay of a rapidly moving virtual photon. The amplitude for the emission of a soft gluon was given in Eq. (58). Squaring, summing over colours and including the gluon phase-space we get the following result:

$$\begin{aligned} d\sigma_g &= \sum |A_{\text{soft}}|^2 \frac{d^3k}{(2\pi)^3 2k^0} \sum |A_0|^2 \frac{-2p^\mu \bar{p}^\nu}{(pk)(\bar{p}k)}\, g^2 \sum \epsilon_\mu \epsilon^*_\nu \frac{d^3k}{(2\pi)^3 2k^0} \\ &= d\sigma_0 \frac{2(p\bar{p})}{(pk)(\bar{p}k)}\, g^2\, C_{\text{F}} \left(\frac{d\phi}{2\pi}\right) \frac{k^0 dk^0}{8\pi^2}\, d\,\cos\theta \\ &= d\sigma_0 \frac{\alpha_s C_{\text{F}}}{\pi} \frac{dk^0}{k^0} \frac{d\phi}{2\pi} \frac{1-\cos\theta_{ij}}{(1-\cos\theta_{ik})(1-\cos\theta_{jk})}\, d\cos\theta \end{aligned} \tag{66}$$

where $\theta_{\alpha\beta} = \theta_\alpha - \theta_\beta$, and i, j, k refer to the q, $\bar{q}$ and gluon directions, respectively. We can write the following identity:

$$\frac{1-\cos\theta_{ij}}{(1-\cos\theta_{ik})(1-\cos\theta_{jk})} = \frac{1}{2}\left[\frac{\cos\theta_{jk}-\cos\theta_{ij}}{(1-\cos\theta_{ik})(1-\cos\theta_{jk})} + \frac{1}{1-\cos\theta_{ik}}\right] + \frac{1}{2}[i \leftrightarrow j] \equiv W_{(i)} + W_{(j)}\,. \tag{67}$$

We would like to interpret the two functions $W_{(i)}$ and $W_{(j)}$ as radiation probabilities from the quark and antiquark lines. Each of them is in fact only singular in the limit of gluon emission parallel to the respective quark:

$$W_{(i)} \;\rightarrow\; \text{finite if } k \parallel j \;\; (\cos\theta_{jk} \to 1) \tag{68}$$

$$W_{(j)} \;\rightarrow\; \text{finite if } k \parallel i \;\; (\cos\theta_{ik} \to 1)\,. \tag{69}$$

The interpretation as probabilities is however limited by the fact that neither $W_{(i)}$ nor $W_{(j)}$ are positive definite. However, you can easily prove that

$$\int \frac{d\phi}{2\pi} W_{(i)} = \begin{cases} \frac{1}{1-\cos\theta_{ik}} & \text{if } \theta_{ik} < \theta_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{70}$$

where the integral is the azimuthal average around the q direction. A similar result holds for $W_{(j)}$:

$$\int \frac{d\phi}{2\pi} W_{(j)} = \begin{cases} \frac{1}{1-\cos\theta_{jk}} & \text{if } \theta_{jk} < \theta_{ij} \\ 0 & \text{otherwise} \end{cases}\,. \tag{71}$$

As a result, the emission of soft gluons outside the two cones obtained by rotating the antiquark direction around the quark's, and vice-versa, averages to 0. Inside the two cones, one can consider the radiation from the emitters as being uncorrelated. In other words, the two colour lines defined by the quark and antiquark currents act as independent emitters, and the quantum coherence (i.e., the effects of interference between the two graphs contributing to the gluon-emission amplitude) is accounted for by constraining the emission to take place within those fixed cones.

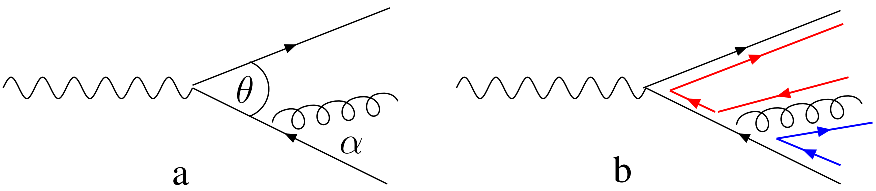


Fig. 13: Radiation off $q\bar{q}$ pair produced by an off-shell photon

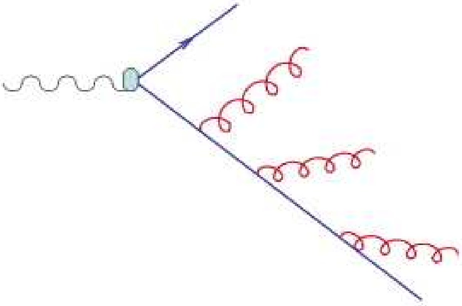


Fig. 14: Collimation of soft gluon emission during the jet evolution

A simple derivation of angular ordering, which more directly exhibits its physical origin, can be obtained as follows. Consider Fig. 13a, which shows a Feynman diagram for the emission of a gluon from a quark line. The quark momentum is denoted by l and the gluon momentum by k; θ is the opening angle between the quark and antiquark, and α is the angle between the nearest quark and the emitted gluon. We will work in the double-log enhanced soft $k^0 << l^0$ and collinear $\alpha << 1$ region. The internal quark propagator $p = (l + k)$ is off-shell, setting the time scale for the gluon emission:

$$\Delta t \simeq \frac{1}{\Delta E} = \frac{l^0}{(k+l)^2} \quad \rightarrow \quad \Delta t \simeq \frac{1}{k^0 \alpha^2} \, . \tag{72}$$

In order to resolve the quarks, the transverse wavelength of the gluon $\lambda_\perp = 1/E_\perp$ must be smaller than the separation between the quarks $b(t) \simeq \theta \, \Delta t$, giving the constraint $1/(\alpha k^0) < \theta \, \Delta t$. Using the results of Eq. (72) for Δt, we arrive at the angular ordering constraint $\alpha < \theta$. Gluon emissions at an angle smaller than θ can resolve the two individual colour quarks and are allowed; emissions at greater angles do not see the colour charge and are therefore suppressed. In processes involving more partons, the angle θ is defined not by the nearest parton, but by the colour connected parton (e.g., the parton that forms a colour singlet with the emitting parton). Figure 13b shows the colour connections for the $q\bar{q}$ event after the gluon is emitted. Colour lines begin on quarks and end on antiquarks. Because gluons are colour octets, they contain the beginning of one line and the end of another.

If one now repeats the exercise for emission of one additional gluon, one will find the same angular constraint, but this time applied to the colour lines defined by the previously established *antenna*. As shown in the previous subsection, the $q\bar{q}g$ state can be decomposed at the leading order in $1/N$ into two independent emitters: one given by the colour line flowing from the gluon to the quark, the other given by the colour line flowing from the antiquark to the gluon. So the emission of the additional gluon will be constrained to take place either within the cone formed by the quark and the gluon, or within the cone formed by the gluon and the antiquark. Either way, the emission angle will be smaller than the angle of the first gluon emission. This leads to the concept of angular ordering, with successive emission of soft gluons taking place within cones which get smaller and smaller, as in Fig. 14.

The fact that colour always flows directly from the emitting parton to the emitted one, the colli-

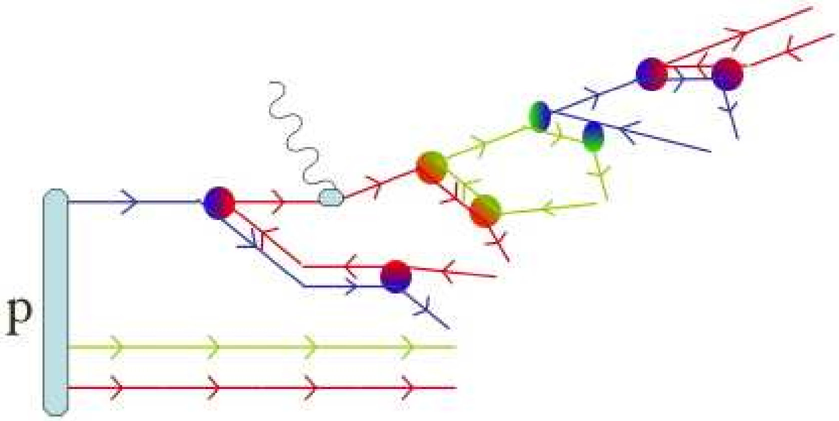


Fig. 15: The colour flow diagram for a deep-inelastic scattering event

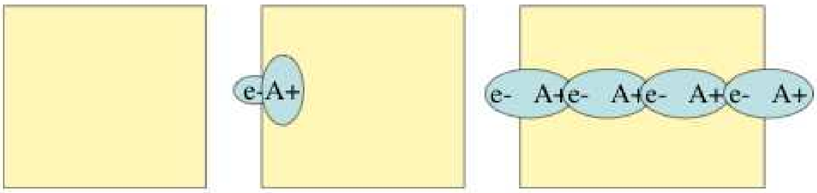


Fig. 16: Charge transfer in a dielectric medium, via a sequence of local polarizations

mation of the jet, and the softening of the radiation emitted at later stages ensure that partons forming a colour-singlet cluster are close in phase-space. As a result, hadronization (the non-perturbative process that will bind together colour-singlet parton pairs) takes place locally inside the jet and is not a long-distance phenomenon connecting partons far away in the evolution tree: only pairs of nearby partons are involved. In particular, there is no direct link between the precise nature of the hard process and the hadronization. These two phases are totally decoupled and, as in the case of the partonic densities, one can infer that hadronization factorizes from the hard process and can be described in a universal (i.e., hard-process independent) fashion. The inclusive properties of jets (particle multiplicity, jet mass, jet broadening, etc.) are independent of the hadronization model, up to corrections of order $(\Lambda/\sqrt{s})^n$ (for some integer power n, which depends on the observable), with $\Lambda \lesssim 1$ GeV.

The final picture, in the case of a DIS event, appears therefore as in Fig. 15. After being deflected by the photon, the struck quark emits the first gluon, which takes away the quark colour and passes on its own anticolour to the escaping quark. This gluon is therefore colour-connected with the last gluon emitted before the hard interaction. As the final-state quark continues its evolution, more and more gluons are emitted, each time leaving their colour behind and transmitting their anticolour to the emerging quark. Angular ordering forces all these gluons to be close in phase-space, until the evolution is stopped once the virtuality of the quark becomes of the order of the strong-interaction scale. The colour of the quark is left behind, and when hadronization takes over it is only the nearby colour-connected gluons which are transformed, with a phenomenological model, in hadrons. This mechanism for the transfer of colour across subsequent gluon emissions is similar to what happens when we place a charge near the surface of a dielectric medium. This will become polarized, and a charge will appear on the opposite end of the medium. The appearance of the charge is the result of a sequence of local charge shifts, whereby neighbouring atoms get polarized, as in Fig. 16.

5 Applications to hadronic collisions

In hadronic collisions, all phenomena are QCD-related. The dynamics is more complex than in e^+e^- or DIS, since both beam and target have a non-trivial partonic structure. As a result, calculations (and experimental analyses) are more complicated. QCD phenomenology is however much richer, and the higher energies available in hadronic collisions allow us to probe the structure of the proton and of its constituents at the smallest scales attainable in a laboratory.

Contrary to the case of e^+e^- and lepton–hadron collisions, where calculations are routinely available up to next-to-next-to-leading order (NNLO) accuracy, theoretical calculations for hadronic collisions are available at best with next-to-leading-order (NLO) accuracy. The only exception is the case of Drell–Yan production, where NNLO results are known for the total cross-sections. So we generally have relatively small precision in the theoretical predictions, and theoretical uncertainties which are large when compared to LEP or HERA.

However, $p\bar{p}$ collider physics is primarily *discovery* physics, rather than precision physics. (There are exceptions, such as the measurements of the W mass and of the properties of b-hadrons; but these are not QCD-related measurements.) As such, knowledge of QCD is essential both for the estimate of the expected signals, and for the evaluation of the backgrounds. Tests of QCD in $p\bar{p}$ collisions confirm our understanding of perturbation theory, or, when they fail, point to areas where our approximations need to be improved. (see, e.g., the theory advances prompted by the measurements of ψ production by CDF at the Tevatron!).

Finally, a reliable theoretical control over the details of production dynamics allows one to extract important information on the structure of the proton (parton densities) in regions of Q^2 and x otherwise inaccessible. Control of QCD at the current machines (the Tevatron at Fermilab) is therefore essential for the extrapolation of predictions to higher energies (say for applications at the future LHC, at CERN).

The key ingredients for the calculation of production rates and distributions in hadronic collisions are

- the matrix elements for the hard, partonic process (e.g., $gg \to gg, gg \to b\bar{b}, q\bar{q}' \to W, \ldots$);
- the hadronic parton densities, discussed in the previous lecture.

Then the production rate for a given final state H is given by a factorization formula similar to the one used to describe DIS:

$$d\sigma(p\bar{p} \to H + X) = \int dx_1\, dx_2 \sum_{i,j} f_i(x_1, Q)\, f_j(x_2.Q)\, d\hat{\sigma}(ij \to H) \qquad (73)$$

where the parton densities f_i are evaluated at a scale Q typical of the hard process under consideration. For example $Q \simeq M_{\mathrm{DY}}$ for production of a Drell–Yan pair, $Q \simeq E_{\mathrm{T}}$ for high transverse-energy (E_{T}) jets, $Q^2 \simeq p_{\mathrm{T}}^2 + m_Q^2$ for high-p_{T} heavy quarks, etc.

In this lecture we will briefly explore two of the QCD phenomena currently studied in hadronic collisions: Drell–Yan, and inclusive jet production. More details can be found in Refs. [4, 8].

5.1 Drell–Yan processes

While the Z boson has been recently studied with great precision by the LEP experiments, it was actually discovered, together with the W boson, by the CERN experiments UA1 and UA2 in $p\bar{p}$ collisions. W physics is now being studied in great detail at LEP2, but the best direct measurements of its mass by a single group still belong to $p\bar{p}$ experiments (CDF and D0 at the Tevatron). Even after the ultimate luminosity will have been accumulated at LEP2, with a great improvement in the determination of the parameters of the W boson, the monopoly of W studies will immediately return to hadron colliders, with the Tevatron data-taking resuming in the year 2000, and later on with the start of the LHC experiments.

Precision measurements of W production in hadronic collisions are important for several reasons:

- this is the only process in hadronic collisions which is known to NNLO accuracy;
- the rapidity distribution of the charged leptons from W decays is sensitive to the ratio of the up- and down-quark densities, and can contribute to our understanding of the proton structure;
- deviations from the expected production rates of highly virtual W's ($p\bar{p} \to W^* \to e\nu$) are a possible signal of the existence of new W bosons, and therefore of new gauge interactions.

The partonic cross-section for the production of a W boson from the annihilation of a $q\bar{q}$ pair can be easily calculated, giving the following result [4, 8]:

$$\hat{\sigma}(q_i\bar{q}_j \to W) \;=\; \pi\,\frac{\sqrt{2}}{3}\,|V_{ij}|^2\,G_{\rm F}\,M_W^2\,\delta(\hat{s}-M_W^2) \;=\; A_{ij}\,M_W^2\,\delta(\hat{s}-M_W^2) \tag{74}$$

where $\hat{s}$ is partonic centre-of-mass (c.m.) energy squared, and V_{ij} is the element of the Cabibbo–Kobayashi–Maskawa matrix. The delta function comes from the $2 \to 1$ phase-space, which forces the c.m. energy of the initial state to coincide with the W mass. It is useful to introduce the two variables

$$\tau \;=\; \frac{\hat{s}}{S_{\rm had}} \equiv x_1 x_2 \tag{75}$$

$$y \;=\; \frac{1}{2}\log\left(\frac{E_W+p_W^z}{E_W-p_W^z}\right) \equiv \frac{1}{2}\log\left(\frac{x_1}{x_2}\right)\,, \tag{76}$$

where $S_{\rm had}$ is the hadronic c.m. energy squared. The variable y is called *rapidity*. For slowly moving objects it reduces to the standard velocity, but, unlike the velocity, it transforms additively even at high energies under Lorentz boosts along the direction of motion. Written in terms of τ and y, the integration measure over the initial-state parton momenta becomes $dx_1dx_2 = d\tau dy$. Using this expression and Eq. (74) in Eq. (73), we obtain the following result for the leading-order total W production cross-section:

$$\sigma_{\rm DY} \;=\; \sum_{i,j}\frac{\pi\,A_{ij}}{M_W^2}\,\tau\int_\tau^1\frac{dx}{x}\,f_i(x)\,f_j\left(\frac{\tau}{x}\right) \equiv \sum_{i,j}\frac{\pi\,A_{ij}}{M_W^2}\tau\mathcal{L}_{ij}(\tau) \tag{77}$$

where the function $\mathcal{L}_{ij}(\tau)$ is usually called *partonic luminosity*. In the case of $u\bar{d}$ collisions, the overall factor in front of this expression has a value of approximately 6.5 nb. It is interesting to study the partonic luminosity as a function of the hadronic c.m. energy. This can be done by taking a simple approximation for the parton densities. Following the indications of the figures presented in the previous lecture, we shall assume that $f_i(x) \sim 1/x^{1+\delta}$, with $\delta < 1$. Then

$$\mathcal{L}(\tau) \;=\; \int_\tau^1\frac{dx}{x}\,\frac{1}{x^{1+\delta}}\left(\frac{x}{\tau}\right)^{1+\delta} = \frac{1}{\tau^{1+\delta}}\int_\tau^1\frac{dx}{x} \;=\; \frac{1}{\tau^{1+\delta}}\,\log\left(\frac{1}{\tau}\right) \tag{78}$$

and

$$\sigma_W \sim \tau^{-\delta}\,\log\left(\frac{1}{\tau}\right) = \left(\frac{S_{\rm had}}{M_W^2}\right)^\delta\log\left(\frac{S_{\rm had}}{M_W^2}\right)\,. \tag{79}$$

The Drell-Yan cross-section grows therefore at least logarithmically with the hadronic c.m. energy. This is to be compared with the behaviour of the Z production cross-section in e^+e^- collisions, which is steeply diminishing for values of s well above the production threshold. The reason for the different behaviour in hadronic collisions is that while the energy of the hadronic initial state grows, it will always be possible to find partons inside the hadrons with the appropriate energy to produce the W directly on-shell. The number of partons available for the production of a W is furthermore increasing with the increase in hadronic energy, since the larger the hadron energy, the smaller will be the value of hadron

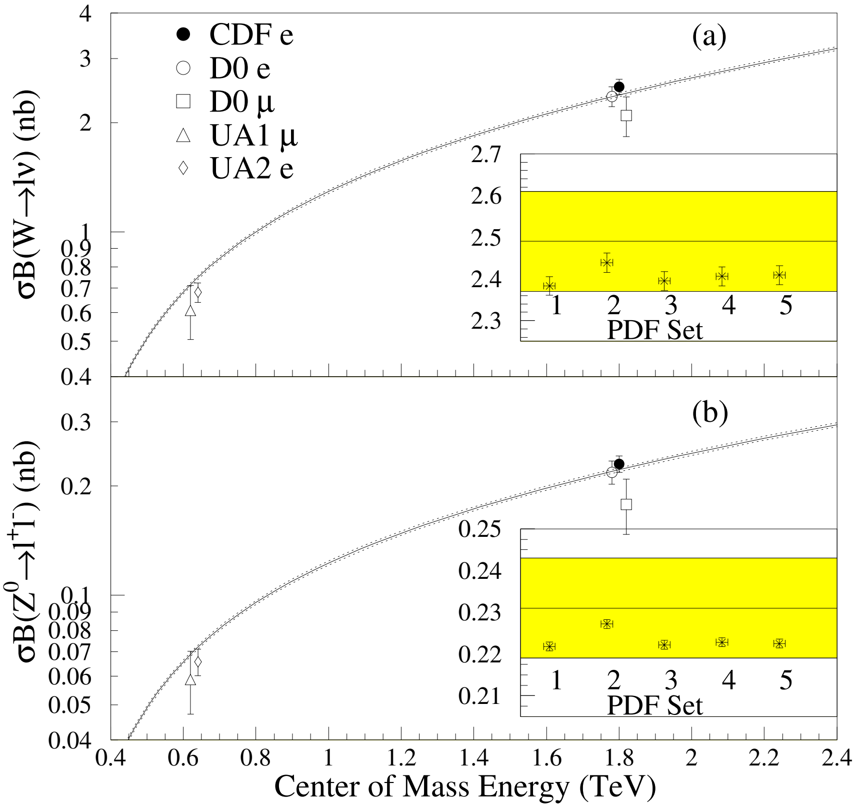


Fig. 17: Comparison of measured (a) $\sigma{\cdot}B(W \to e\nu)$ and (b) $\sigma{\cdot}B(Z^0 \to e^+e^-)$ to two-loop theoretical predictions using MRSA parton distribution functions. The UA1 and UA2 measurements and D0 measurements are offset horizontally by ± 0.02 TeV for clarity. In the inset, the shaded area shows the 1σ region of the CDF measurement; the stars show the predictions using various parton distribution function sets (1) MRSA, (2) MRSD0′, (3) MRSD-′, (4) MRSH and (5) CTEQ2M. The theoretical points include a common uncertainty in the predictions from choice of renormalization scale ($M_W/2$ to $2M_W$).

momentum fraction x necessary to produce the W. The increasing number of partons available at smaller and smaller values of x causes then the growth of the total W production cross-section.

A comparison between the best available prediction for the production rates of W and Z bosons in hadronic collisions, and the experimental data, is shown in Fig. 17. The experimental uncertainties will soon be dominated by the limited knowledge of the machine luminosity, and will exceed the accuracy of the NNLO predictions. This suggests that in the future the total rate of produced W bosons could be used as an accurate luminometer.

It is also interesting to note that an accurate measurement of the relative W and Z production rates (which is not affected by the knowledge of the total integrated luminosity, which will cancel in their ratio) provides a tool to measure the total W width. This can be seen from the following equation:

$$\Gamma_W \;=\; \underset{\substack{\uparrow\\ \text{measure}}}{\frac{N^{\text{obs}}(Z \to e^+e^-)}{N^{\text{obs}}(W \to e^\pm\nu)}} \; \underset{\substack{\nwarrow \quad\quad \nearrow\\ \text{calculable}}}{\left(\frac{\sigma_{W^\pm}}{\sigma_Z}\right) \left(\frac{\Gamma^W_{e\nu}}{\Gamma^Z_{e^+e^-}}\right)} \; \underset{\substack{\uparrow\\ \text{LEP/SLC}\,.}}{\Gamma_Z}$$

As of today, this technique provides the best measurement of Γ_W: $\Gamma_W = 2.06 \pm 0.06$ GeV, which is a factor of 5 more accurate than the current best direct measurements from LEP2.

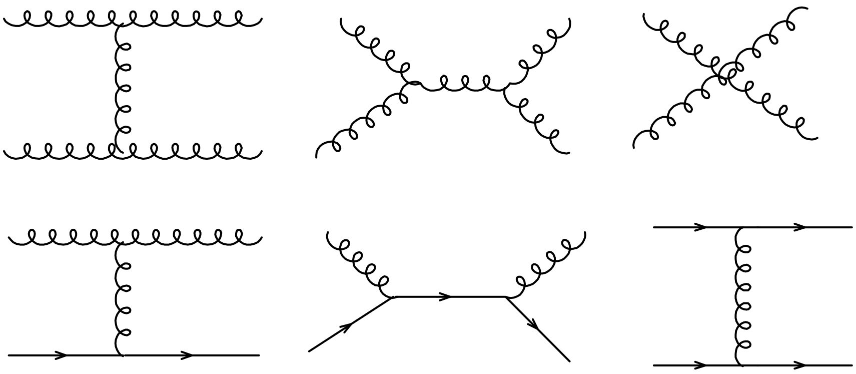


Fig. 18: Representative diagrams for the production of jet pairs in hadronic collisions

5.2 *W* rapidity asymmetry

The measurement of the charge asymmetry in the rapidity distribution of W bosons produced in $p\bar{p}$ collisions can provide an important measurement of the ratio of the u-quark and d-quark momentum distributions. Using the formulas provided above, you can in fact easily check as an exercise that

$$\frac{d\sigma_{W^+}}{dy} \propto f_u^p(x_1)\, f_{\bar{d}}^{\bar{p}}(x_2) + f_{\bar{d}}^p(x_1) f_u^{\bar{p}}(x_2) \tag{80}$$

$$\frac{d\sigma_{W^-}}{dy} \propto f_{\bar{u}}^p(x_1)\, f_d^{\bar{p}}(x_2) + f_d^p(x_1) f_{\bar{u}}^{\bar{p}}(x_2)\,. \tag{81}$$

We can then construct the following charge asymmetry (assuming the dominance of the quark densities over the antiquark ones, which is valid in the kinematical region of interest for W production at the Tevatron):

$$A(y) = \frac{\frac{d\sigma_{W^+}}{dy} - \frac{d\sigma_{W^-}}{dy}}{\frac{d\sigma_{W^+}}{dy} + \frac{d\sigma_{W^-}}{dy}} = \frac{f_u^p(x_1)\, f_d^p(x_2) - f_d^p(x_1)\, f_u^p(x_2)}{f_u^p(x_1)\, f_d^p(x_2) + f_d^p(x_1)\, f_u^p(x_2)}\,. \tag{82}$$

Setting $f_d(x) = f_u(x)\; R(x)$, we then get

$$A(y) = \frac{R(x_2) - R(x_1)}{R(x_2) + R(x_1)}\,, \tag{83}$$

which measures the $R(x)$ ratio since $x_{1,2}$ are known in principle from the kinematics: $x_{1,2} = \sqrt{\tau}\exp(\pm y)$.[2] The current CDF data provide the most accurate measurement to date of this quantity (see Ref. [8]).

5.3 Jet production

Jet production is the hard process with the largest rate in hadronic collisions. For example, the cross-section for producing at the Tevatron ($\sqrt{S_{\text{had}}} = 1.8$ TeV) jets of transverse energy $E_T^{\text{jet}} \lesssim 50$ GeV is of the order of a μb. This means 50 events per second at the luminosities available at the Tevatron. The data collected at the Tevatron so far extend all the way up to the E_T values of the order of 450 GeV. These events are generated by collisions among partons which carry over 50% of the available $p\bar{p}$ energy, and allow us to probe the shortest distances ever reached. The leading mechanisms for jet production are shown in Fig. 18.

The two-jet inclusive cross-section can be obtained from the formula

$$d\sigma = \sum_{ijkl} dx_1\, dx_2\, f_i^{(H_1)}(x_1, \mu)\, f_j^{(H_2)}(x_2, \mu)\, \frac{d\hat{\sigma}_{ij \to k+l}}{d\Phi_2}\, d\Phi_2\,, \tag{84}$$

[2] In practice one cannot determine $x_{1,2}$ with arbitrary precision on an event-by-event basis, since the longitudinal momentum of the neutrino cannot be easily measured. The actual measurement is therefore done by studying the charge asymmetry in the rapidity distribution of the charged lepton.

which has to be expressed in terms of the rapidity and transverse momentum of the quarks (or jets) in order to make contact with physical reality. The two-particle phase-space is given by

$$d\Phi_2 = \frac{d^3k}{2k^0(2\pi)^3}\, 2\pi\, \delta\left[(p_1+p_2-k)^2\right] \,, \tag{85}$$

and, in the c.m. of the colliding partons, we get

$$d\Phi_2 = \frac{1}{2(2\pi)^2}\, d^2k_{\mathrm{T}}\, dy\, 2\,\delta\left[\hat{s}-4(k^0)^2\right] \,, \tag{86}$$

where k_{T} is the transverse momentum of the final-state partons. Here y is the rapidity of the produced parton in the parton c.m. frame. It is given by

$$y = \frac{y_1 - y_2}{2} \tag{87}$$

where y_1 and y_2 are the rapidities of the produced partons in the laboratory frame (in fact, in any frame). One also introduces

$$y_0 = \frac{y_1+y_2}{2} = \frac{1}{2}\log\frac{x_1}{x_2} \,, \qquad \tau = \frac{\hat{s}}{S_{\text{had}}} = x_1\, x_2 \,. \tag{88}$$

We have

$$dx_1\, dx_2 = dy_0\, d\tau \,. \tag{89}$$

We obtain

$$d\sigma = \sum_{ijkl} dy_0\, \frac{1}{S_{\text{had}}}\, f_i^{(H_1)}(x_1,\mu)\, f_j^{(H_2)}(x_2,\mu)\, \frac{d\hat{\sigma}_{ij\to k+l}}{d\Phi_2}\, \frac{1}{2(2\pi)^2}\, 2\, dy\, d^2k_{\mathrm{T}} \tag{90}$$

which can also be written as

$$\frac{d\sigma}{dy_1\, dy_2\, d^2k_{\mathrm{T}}} = \frac{1}{S_{\text{had}}\; 2(2\pi)^2} \sum_{ijkl} f_i^{(H_1)}(x_1,\mu)\, f_j^{(H_2)}(x_2,\mu)\, \frac{d\hat{\sigma}_{ij\to k+l}}{d\Phi_2} \,. \tag{91}$$

The variables x_1, x_2 can be obtained from y_1, y_2 and k_{T} from the equations

$$\begin{aligned} y_0 &= \frac{y_1+y_2}{2} && (92)\\ y &= \frac{y_1-y_2}{2} && (93)\\ x_{\mathrm{T}} &= \frac{2k_{\mathrm{T}}}{\sqrt{S_{\text{had}}}} && (94)\\ x_1 &= x_{\mathrm{T}}\, e^{y_0}\, \cosh y && (95)\\ x_2 &= x_{\mathrm{T}}\, e^{-y_0}\, \cosh y \,. && (96) \end{aligned}$$

For the partonic variables, we need $\hat{s}$ and the scattering angle in the parton c.m. frame θ, since

$$t = -\frac{\hat{s}}{2}\,(1-\cos\theta) \,, \qquad u = -\frac{\hat{s}}{2}\,(1+\cos\theta) \,. \tag{97}$$

Neglecting the parton masses, you can show that the rapidity can also be written as

$$y = -\log\tan\frac{\theta}{2} \equiv \eta \,, \tag{98}$$

with η usually being referred to as pseudorapidity.

The leading-order Born cross-sections for parton–parton scattering are reported in Table 1.

Table 1: Cross-sections for light parton scattering. The notation is $p_1\, p_2 \to k\, l$, $\hat{s} = (p_1 + p_2)^2$, $\hat{t} = (p_1 - k)^2$, $\hat{u} = (p_1 - l)^2$.

Process	$\frac{d\hat{\sigma}}{d\Phi_2}$
$qq' \to qq'$	$\frac{1}{2\hat{s}}\frac{4}{9}\frac{\hat{s}^2+\hat{u}^2}{\hat{t}^2}$
$qq \to qq$	$\frac{1}{2}\frac{1}{2\hat{s}}\left[\frac{4}{9}\left(\frac{\hat{s}^2+\hat{u}^2}{\hat{t}^2}+\frac{\hat{s}^2+\hat{t}^2}{\hat{u}^2}\right)-\frac{8}{27}\frac{\hat{s}^2}{\hat{u}\hat{t}}\right]$
$q\bar{q} \to q'\bar{q}'$	$\frac{1}{2\hat{s}}\frac{4}{9}\frac{\hat{t}^2+\hat{u}^2}{\hat{s}^2}$
$q\bar{q} \to q\bar{q}$	$\frac{1}{2\hat{s}}\left[\frac{4}{9}\left(\frac{\hat{s}^2+\hat{u}^2}{\hat{t}^2}+\frac{\hat{t}^2+\hat{u}^2}{\hat{s}^2}\right)-\frac{8}{27}\frac{\hat{u}^2}{\hat{s}\hat{t}}\right]$
$q\bar{q} \to gg$	$\frac{1}{2}\frac{1}{2\hat{s}}\left[\frac{32}{27}\frac{\hat{t}^2+\hat{u}^2}{\hat{t}\hat{u}}-\frac{8}{3}\frac{\hat{t}^2+\hat{u}^2}{\hat{s}^2}\right]$
$gg \to q\bar{q}$	$\frac{1}{2\hat{s}}\left[\frac{1}{6}\frac{\hat{t}^2+\hat{u}^2}{\hat{t}\hat{u}}-\frac{3}{8}\frac{\hat{t}^2+\hat{u}^2}{\hat{s}^2}\right]$
$gq \to gq$	$\frac{1}{2\hat{s}}\left[-\frac{4}{9}\frac{\hat{s}^2+\hat{u}^2}{\hat{s}\hat{u}}+\frac{\hat{u}^2+\hat{s}^2}{\hat{t}^2}\right]$
$gg \to gg$	$\frac{1}{2}\frac{1}{2\hat{s}}\frac{9}{2}\left(3-\frac{\hat{t}\hat{u}}{\hat{s}^2}-\frac{\hat{s}\hat{u}}{\hat{t}^2}-\frac{\hat{s}\hat{t}}{\hat{u}^2}\right)$

It is interesting to note that a good approximation to the exact results can be easily obtained by using the soft-gluon techniques introduced in the third lecture. Based on the fact that even at 90° min($|t|, |u|$) does not exceed $s/2$, and that therefore everything else being equal a propagator in the t or u channel contributes to the square of an amplitude four times more than a propagator in the s channel, it is reasonable to assume that the amplitudes are dominated by the diagrams with a gluon exchanged in the t (or u) channel. It is easy to calculate the amplitudes in this limit using the soft-gluon approximation. For example, the amplitude for the exchange of a soft gluon among a qq' pair is given by

$$(\lambda^a_{ij})\,(\lambda^a_{kl})\,2p_\mu\,\frac{1}{t}\,2p'_\mu \;=\; \lambda^a_{ij}\,\lambda^a_{kl}\,\frac{4p\cdot p'}{t} = \frac{2s}{t}\,\lambda^a_{ij}\,\lambda^a_{kl}\,. \tag{99}$$

The p_μ and p'_μ factors represent the coupling of the exchanged gluon to the q and q' quark lines, respectively [see Eq. (59)]. Squaring, and summing and averaging over spins and colours, gives

$$\overline{\sum_{\text{colours, spin}}}|M_{qq'}|^2 \;=\; \frac{1}{N^2}\left(\frac{N^2-1}{4}\right)\frac{4s^2}{t^2} = \frac{8}{9}\,\frac{s^2}{t^2}\,. \tag{100}$$

Since for this process the diagram with a t-channel gluon exchange is symmetric for $s \leftrightarrow u$ exchange, and since $u \to -s$ in the $t \to 0$ limit, the above result can be rewritten in an explicitly (s,u) symmetric way as

$$\frac{4}{9}\,\frac{s^2+u^2}{t^2} \tag{101}$$

which indeed exactly agrees with the result of the exact calculation, as given in Table 1. The corrections which appear from s or u gluon exchange when the quark flavours are the same or when we study a $q\bar{q}$ process are small, as can be seen by comparing the above result with the expressions in the table.

As another example we consider the case of $qg \to qg$ scattering. The amplitude will be exactly the same as in the $qq' \to qq'$ case, up to the different colour factors. A simple calculation then gives

$$\overline{\sum_{\text{colours, spin}}}|M_{qg}|^2 \;=\; \frac{9}{4}\overline{\sum}|M_{qq'}|^2 = \frac{s^2+u^2}{t^2}\,. \tag{102}$$

The exact result is

$$\frac{u^2+s^2}{t^2} - \frac{4}{9}\,\frac{u^2+s^2}{us} \tag{103}$$

which even at 90°, the point where the t-channel exchange approximation is worst, only differs from this latter by no more than 25%.

As a final example we consider the case of $gg \to gg$ scattering, which in our approximation gives

$$\overline{\sum}|M_{gg}|^2 = \frac{9}{2}\frac{s^2}{t^2} \, . \tag{104}$$

By $u \leftrightarrow t$ symmetry, we should expect the simple improvement

$$\overline{\sum}|M_{gg}|^2 \sim \frac{9}{2}\left(\frac{s^2}{t^2} + \frac{s^2}{u^2}\right) \, . \tag{105}$$

This only differs by 20% from the exact result at 90°.

Note that at small t the following relation holds:

$$\hat{\sigma}_{gg} : \hat{\sigma}_{qg} : \hat{\sigma}_{q\bar{q}} = \left(\frac{9}{4}\right) : 1 : \left(\frac{4}{9}\right) \, . \tag{106}$$

The $9/4$ factors are simply the ratios of the colour factors for the coupling to gluons of a gluon (C_{A}) and of a quark (T_{F}), after including the respective colour-average factors: $(1/(N^2-1)$ for the gluon, and $1/N$ for the quark). Using Eq. (106), we can then write

$$d\sigma_{\text{hadr}} \;=\; \int dx_1\, dx_2 \sum_{i,j} f_i(x_1)\, f_j(x_2)\, d\hat{\sigma}_{ij} \;=\; \int dx_1\, dx_2\, F(x_1)\, F(x_2)\, d\hat{\sigma}_{gg}(gg \to \text{jets}) \tag{107}$$

where the object

$$F(x) = f_g(x) + \frac{4}{9}\sum_f \left[q_f(x) + \bar{q}_f(x)\right] \tag{108}$$

is usually called the *effective structure function*. This result indicates that the measurement of the inclusive jet cross-section does not allow us in principle to disentangle the independent contribution of the various partonic components of the proton, unless of course one is considering a kinematical region where the production is dominated by a single process. The relative contributions of the different channels, as predicted using the global fits of parton densities available in the literature, are shown in Fig. 19.

Predictions for jet production at colliders are available today at next-to-leading order in QCD. A comparison between these calculations and the available data is given in Figs. 20 and 21. At the Tevatron, jets up to 600 GeV transverse momentum have been observed. That is $x \gtrsim 0.6$ and $Q^2 \simeq 400\,000$ GeV^2. This is a domain of x and Q^2 not accessible to HERA. The current agreement between theory and data is excellent over eight orders of magnitude of cross-section, from $E_{\mathrm{T}} \sim 50$ to $E_{\mathrm{T}} \sim 600$ GeV. The experimental and theoretical systematic uncertainties, however, become larger than 30% when $E_{\mathrm{T}} \gtrsim$ 400 GeV, preventing a very accurate test of the smallest scales. More data on jet production at large rapidity will allow us to reduce the PDF uncertainties at large x. The uncertainty in the absolute energy scale remains however a critical and difficult to overcome experimental limitation at the highest energies.

Appendices

A Renormalization, or “Theorists are not afraid of infinities!”

QCD calculations are extremely demanding. Although perturbative, the size of the coupling constant even at rather large values of the exchanged momentum, Q^2, is such that the convergence of the perturbative expansion is slow. Several orders of perturbation theory are required in order to obtain a good accuracy. The complexity of the calculations grows dramatically with the order of the approximation. As

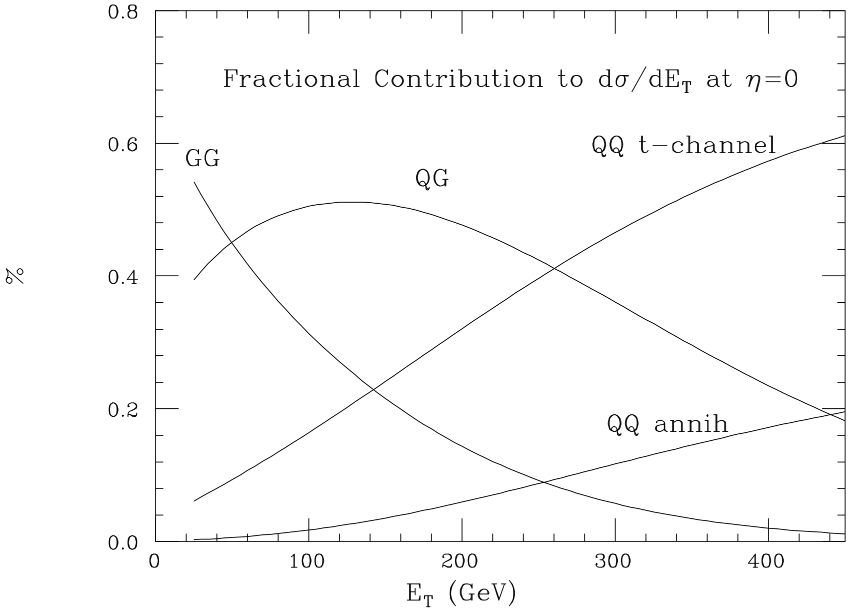


Fig. 19: Relative contribution to the inclusive jet-E_{T} rates from the different production channels

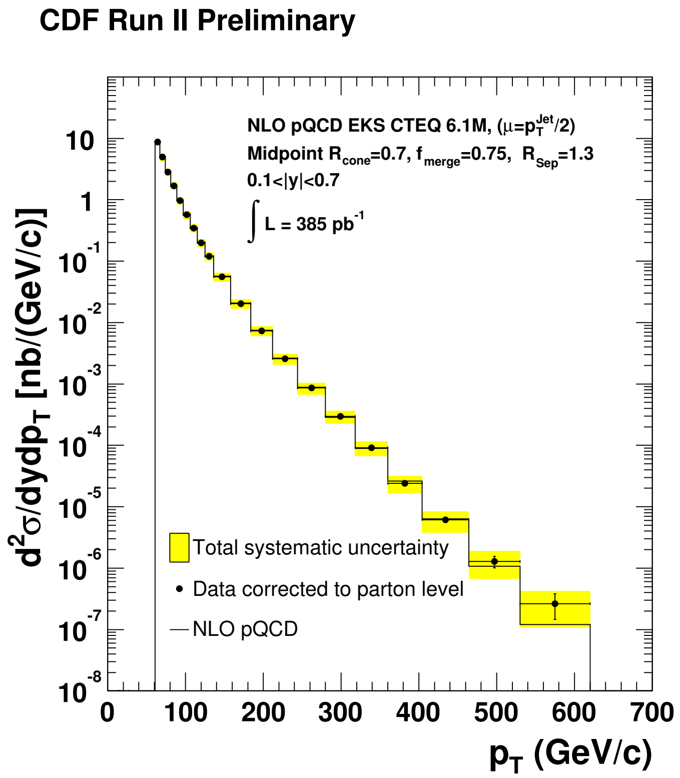


Fig. 20: Inclusive E_{T} spectra for central jets at the Tevatron

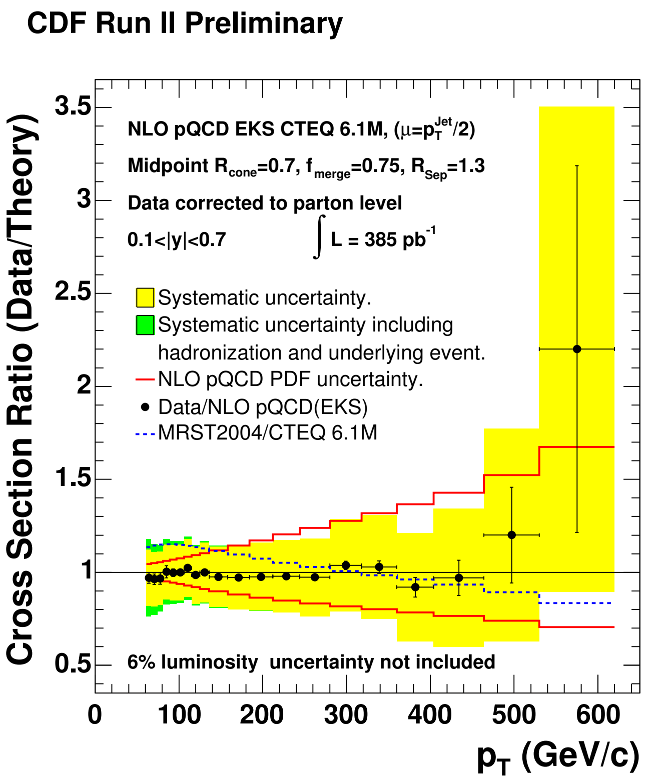
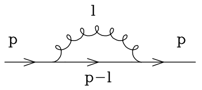


Fig. 21: Comparison of inclusive jet cross-sections with QCD calculations at the Tevatron

an additional complication, the evaluation of a large class of higher-order diagrams gives rise to results which are a priori ill-defined, namely to infinities. A typical example of what is known as an *ultraviolet* divergence appears when considering the corrections to the quark self-energy. Using the Feynman rules presented Section 2, one can obtain



$$= (-ig)^2\, C_{\rm F} \int \frac{d^4\ell}{(2\pi)^4}\gamma_\mu \frac{i}{\not p+\ell}\gamma_\nu \left(-\frac{ig^{\mu\nu}}{\ell^2}\right) \equiv i\not p\Sigma(p)\,, \tag{A.1}$$

where simple manipulations lead to the following expression for $\Sigma(p)$:

$$\Sigma(p) = iC_{\rm F}\int \frac{d^4\ell}{(2\pi)^4}\,\frac{1}{\ell^2(p+\ell)^2}\,, \tag{A.2}$$

which is logarithmically divergent in the ultraviolet ($|\ell| \to \infty$) region. In this appendix we discuss how to deal with these infinities. To start with, we study a simple example taken from standard electrostatics.

A.1 The potential of an infinite line of charge

Let us consider a wire of infinite length, carrying a constant charge density λ. By definition, the dimensions of λ are [length]$^{-1}$. Our goal is to evaluate the electric potential, and eventually the electric field, in a point P at distance R from the wire. There is no need to do any calculation to anticipate that the evaluation of the electric potential will cause some problem. Using the fact that the potential should be linear in the charge density λ, we write $V(R) = \lambda f(R)$. Since the potential itself has the dimensions of [length]$^{-1}$, we clearly see that there is no room for $f(R)$ to have any non-trivial functional dependence on R. The problem is made explicit if we try to evaluate $V(R)$ using the standard electromagnetic formulas:

$$V(R) \;=\; \int \frac{\lambda(r)}{r} dx \;=\; \lambda \int_{-\infty}^{+\infty} \frac{dx}{\sqrt{R^2+x^2}} \tag{A.3}$$

where the integral runs over the position x on the wire. This integral is logarithmically divergent, and the potential is ill-defined. We know however that this is not a serious issue, since the potential itself is not a physical observable, only the electric field is measurable. Since the electric field is obtained by taking the gradient of the scalar potential, it will be proportional to

$$V'(R) \sim \lambda \int_{-\infty}^{+\infty} \frac{dx}{(R^2+x^2)^{3/2}} \,, \tag{A.4}$$

which is perfectly convergent. It is however interesting to explore the possibility of providing a useful operative meaning to the definition of the scalar potential. To do that, we start by *regularizing* the integral in Eq. (A.3). This can be done by introducing the regularized $V(R)$ defined as

$$V_\Lambda(R) \;=\; \int_{-\Lambda}^{\Lambda} \lambda \frac{dx}{\sqrt{R^2+x^2}} \;=\; \lambda \log\left[\frac{\sqrt{\Lambda^2+R^2}+\Lambda}{\sqrt{\Lambda^2+R^2}-\Lambda}\right] \,. \tag{A.5}$$

We can then define the electric field as

$$\vec{E}(R) \;=\; \lim_{\Lambda\to\infty} [-\vec{\nabla} V_\Lambda(R)] \,.$$

It is easy to check that this prescription leads to the right result:

$$\vec{E}(R) \;=\; \lim_{\Lambda\to\infty} \hat{R}\,\frac{2\lambda}{R}\,\frac{\Lambda}{\sqrt{\Lambda^2+R^2}} \to \frac{2\lambda}{R}\,\hat{R} \,. \tag{A.6}$$

Note that in this process we had to introduce a new variable Λ with the dimension of a length. This allows us to solve the puzzle first pointed out at the beginning. At the end, however, the dependence of the physical observable (i.e., the electric field) on this extra parameter disappears. Note also that the object,

$$\delta V = \lim_{\Lambda\to\infty} [V_\Lambda(r_2) - V_\Lambda(r_1)] = \lambda \, \log\left(\frac{r_1^2}{r_2^2}\right) \,, \tag{A.7}$$

is well defined. This suggests a way of defining the potential which is meaningful even in the $\Lambda \to \infty$ limit. We can *renormalize* the potential by subtracting $V(R)$ at some fixed value of $R = R_0$ and taking the $\Lambda \to \infty$ limit:

$$V(R) \to V(R) - V(R_0) = \lambda \log\left(\frac{R_0^2}{R^2}\right) \,. \tag{A.8}$$

The non-physical infinities present in $V(R)$ and $V(R_0)$ cancel each other, leaving a finite result, with a non-trivial R-dependence. Once again, this is possible because a dimensionful parameter (in this case R_0) has been introduced.

This example suggests a strategy for dealing with divergences.

i) Identify an appropriate way to *regularize* infinite integrals.

ii) Absorb the divergent terms into a redefinition of fields or parameters, e.g., via *subtractions*. This step is usually called *renormalization*.

iii) Make sure the procedure is *consistent* by checking that the physical results do not depend on the regularization prescription.

In the rest of this appendix I will explain how this strategy is applied to the case of ultraviolet divergences encountered in perturbation theory.

A.2 Dimensional regularization

The typical expressions we have to deal with have the form

$$I(M^2) = \int \frac{d^4\ell}{(2\pi)^4} \frac{1}{[\ell^2 + M^2]^2} \; . \tag{A.9}$$

It is easy to show that the integral encountered in the quark self-energy diagram can be rewritten as

$$\frac{1}{\ell^2} \frac{1}{(\ell - p)^2} = \int_0^1 dx \, \frac{1}{(L^2 + M^2)^2}, \text{ with } L = \ell - xp, M^2 = x(1-x)p^2 \; . \tag{A.10}$$

The most straightforward extension of the ideas presented above in the case of the infinite charged wire is to regularize the integral using a momentum cutoff, and to renormalize it with a subtraction [for example $I(M^2) - I(M_0^2)$]. Experience has shown, however, that the best way to regularize $I(M^2)$ is to take the analytic continuation of the integral in the number of space-time dimensions. In fact

$$I_D(M^2) = \int \frac{d^D\ell}{(2\pi)^D} \frac{1}{(\ell^2 + M^2)^2} \tag{A.11}$$

is finite $\forall D < 4$. If we could assign a formal meaning to $I_D(M^2)$ for *continuous* values of D away from $D = 4$, we could then perform all our manipulations in $D \neq 4$, regulate the divergences, renormalize fields and couplings, and then go back to $D = 4$.

To proceed, one defines (for Euclidean metrics)

$$d^D\ell = d\Omega_{D-1} \;\; \ell^{D-1} d\ell \tag{A.12}$$

with $d\Omega_{D-1}$ the differential solid angle in D dimensions. Ω_{D-1} is the surface of a D-dimensional sphere. It can be obtained by using the following formal identity:

$$\int d^D\ell \, e^{-\vec{\ell}^2} \equiv \left[\int d\ell \, e^{-\ell^2} \right]^D = \pi^{D/2} \; . \tag{A.13}$$

The integral can also be evaluated, using Eq. (A.12), as

$$\begin{aligned} \int d^D\ell \, e^{-\vec{\ell}^2} &= \Omega_D \int_0^\infty \ell^{D-1} \, e^{-\ell^2} d\ell \; = \; \Omega_D \, \frac{1}{2} \int_0^\infty d\ell^2 (\ell^2)^{\frac{D-2}{2}} \, e^{-\ell^2} \\ &= \Omega_D \, \frac{1}{2} \int_0^\infty dx \, e^{-x} x^{\frac{D-2}{2}} \equiv \frac{\Omega_D}{2} \Gamma\left(\frac{D}{2}\right) \; . \end{aligned} \tag{A.14}$$

Comparing Eqs. (A.13) and (A.14), we get

$$I_D(M^2) \; = \; \frac{1}{(4\pi)^{D/2}} \frac{1}{\Gamma(D/2)} \int_0^\infty dx \, x^{\frac{D-2}{2}} (x{+}M^2)^2 \; = \; \frac{1}{(4\pi)^{D/2}} \frac{\Gamma(2 - D/2)}{\Gamma(2)} (M^2)^{\frac{D}{2}-2} \, . \tag{A.15}$$

Defining $D = 4 - 2\epsilon$ (with the understanding that ϵ will be taken to 0 at the end of the day), and using the small-ϵ expansion,

$$\Gamma(\epsilon) = \frac{1}{\epsilon} - \gamma_\epsilon + \mathcal{O}(\epsilon) \; , \tag{A.16}$$

we finally obtain

$$(4\pi)^2 I_D(M^2) \to \frac{1}{\epsilon} - \log 4\pi M^2 - \gamma_\epsilon \; . \tag{A.17}$$

The divergent part of the integral is then regularized as a pole in $(D - 4)$. The M-dependent part of the integral behaves logarithmically, as expected because the integral itself was dimensionless in $D = 4$. The $1/\epsilon$ pole can be removed by a subtraction:

$$I(M^2) = I(\mu^2) + (4\pi)^2 \log\left(\frac{\mu^2}{M^2}\right) \; , \tag{A.18}$$

where the subtraction scale μ^2 is usually referred to as the ‘renormalization scale’.

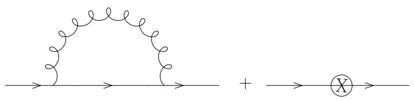
One can prove (and you will find this in the quoted textbooks) that other divergent integrals which appear in other loop diagrams can be regularized in a similar fashion, with the appearance of $1/\epsilon$ poles. Explicit calculations and more details on this technique can be found in the bibliography.

A.3 Renormalization

Let us come back now to our quark self-energy diagram, Eq. (A.1). After regulating the divergence using dimensional regularization, we can eliminate it by adding a counterterm to the Lagrangian:

$$\mathcal{L} \to \mathcal{L} + \Sigma(p)\bar{\psi} i \partial\!\!\!/ \psi \;=\; [1+\Sigma(p)]\bar{\psi} i \partial\!\!\!/ \psi + \dots \; . \tag{A.19}$$

In this way, the corrections at $O(g^2)$ to the inverse propagator are finite:



$$= -i p\!\!\!/ \Sigma(p) + i p\!\!\!/ \Sigma(p) = 0 \; . \tag{A.20}$$

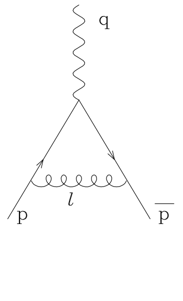
The inclusion of this counterterm can be interpreted as a renormalization of the quark wave function. To see this, it is sufficient to define

$$\psi_{\mathrm{R}} \;=\; \left[1+\Sigma(p^2)\right]^{1/2} \psi \tag{A.21}$$

and verify that the kinetic part of the Lagrangian written in terms of ψ_{R} takes again the canonical form.

It may seem that this regularization/renormalization procedure can always be carried out, with all possible infinities being removed by *ad hoc* counter-terms. This is not true. That these subtractions can be performed consistently for any possible type of divergence which develops in perturbation theory is a highly non-trivial fact. To convince you of this, consider the following example.

Let us study the QCD corrections to the interaction of quarks with a photon:



$$\begin{aligned}
&= (-ig)^2 C_{\mathrm{F}} \int \frac{d^4\ell}{(2\pi)^4} \left[\gamma^p \, \frac{i}{p\!\!\!/ + \ell\!\!\!/} \overbrace{(-i\, e\gamma^\mu)}^{\Gamma^\mu} \, \frac{i}{\bar{p}\!\!\!/ + \ell\!\!\!/} \, \gamma^p \right] \left(\frac{-i}{\ell^2} \right) \\
&= -ig^2 C_{\mathrm{F}} \int \frac{d^4\ell}{(2\pi)^4} \, (-2)(\bar{p}\!\!\!/ + \ell\!\!\!/)\Gamma^\mu(p\!\!\!/ + \ell\!\!\!/) \, \frac{1}{\ell^2(p+\ell)^2(\bar{p}+\ell)^2} \\
&\xrightarrow{\text{leading div.}} -ig^2(-2)C_{\mathrm{F}} \int \frac{d^4\ell}{(2\pi)^4} \, \frac{\ell\!\!\!/ \Gamma^\mu \ell\!\!\!/}{\ell^2(p+\ell)^2(\bar{p}+\ell)^2} \overset{\text{def}}{=} ie\gamma^\mu V(q^2) \; .
\end{aligned}$$

It is easily recognized that $V(q^2)$ is divergent. The divergence can be removed by adding a counter-term to the bare Lagrangian:

$$\begin{aligned}
\mathcal{L}_{\text{int}} \;&=\; -e\, A_\mu \bar{\psi}\gamma^\mu\psi \;\to\; -eA_\mu\bar{\psi}\gamma^\mu\psi - eV(q^2)A_\mu\bar{\psi}\gamma^\mu\psi \\
&=\; -[1+V(q^2)]\, e\, A_\mu\bar{\psi}\gamma^\mu\psi \; .
\end{aligned} \tag{A.22}$$

If we take into account the counter-term that was introduced to renormalize the quark self-energy, the part of the quark Lagrangian describing the interaction with photons is now

$$\mathcal{L}_{q,\gamma} = \left[1+\Sigma(p^2)\right] \bar{\psi}\, i\, \partial\!\!\!/ \, \psi - \left[1+V(q^2)\right] \, e\, A_\mu \, \bar{\psi}\gamma^\mu\psi \; . \tag{A.23}$$

Defining a renormalized charge by

$$e_{\mathrm{R}} = e\, \frac{1+V(p^2)}{1+\Sigma(q^2)} \; , \tag{A.24}$$

we are left with the renormalized Lagrangian

$$\mathcal{L}_{\rm R} \;=\; \bar{\psi}_{\rm R}\, i\, \partial\!\!\!/\, \psi_{\rm R} \;+\; e_{\rm R}\, A_\mu\, \bar{\psi}_{\rm R}\, \gamma^\mu\, \psi_{\rm R}\;. \tag{A.25}$$

Can we blindly accept this result, regardless of the values of the counter-terms $V(p^2)$ and $\Sigma(q^2)$? The answer to this question is NO! Charge conservation, in fact, requires $e_{\rm R} = e$. The electric charge carried by a quark cannot be affected by the QCD corrections, and cannot be affected by the renormalization of QCD-induced divergences. There are many ways to see that if $e_{\rm R} \neq e$ the electric charge would not be conserved in strong interactions. The simplest way is to consider the process $e^+\nu_e \to W^+ \to u\bar{d}$. The electric charge of the initial state is +1 in units of e. After including QCD corrections (which in the case of the interaction with a W are the same as those for the interaction of quarks with a photon), the charge of the final state is +1 in units of $e_{\rm R}$. Unless $e_{\rm R} = e$, the total electric charge would not be conserved in this process! It is the non-renormalization of the electric charge in the presence of strong interactions that makes the charge of the proton equal to the sum of the charges of its constituent quarks, in spite of the complex QCD dynamics that holds the quarks together.

As a result, the renormalization procedure is consistent with charge conservation if and only if

$$\frac{V(q^2)}{\Sigma(p^2)} \;\overset{q^2\to 0}{=}\; 1\;. \tag{A.26}$$

This identity should hold at all orders of perturbation theory. It represents a fundamental constraint on the consistency of the theory, and shows that the removal of infinities, by itself, is not a trivial trick which can be applied to arbitrary theories. Fortunately, the previous identity can be shown to hold. You can prove it explicitly at the one-loop order by explicitly evaluating the integrals defining $V(q)$ and $\Sigma(p)$.

To carry out the renormalization programme for QCD at one-loop order, several other diagrams in addition to the quark self-energy need to be evaluated. One needs the corrections to the gluon self-energy, to the coupling of a quark pair to a gluon, and to the three-gluon coupling. Each of these corrections gives rise to infinities, which can be regulated in dimensional regularization. For the purposes of renormalization, it is useful to apply the concept of D dimensions not only to the evaluation of the infinite integrals, but to the full theory as well. In other words, we should consider the Lagrangian as describing the interactions of fields in D dimensions. Nothing changes in its form, but the canonical dimensions of fields and couplings will be shifted. This is because the action (defined as the integral over space-time of the Lagrangian) is a dimensionless quantity. As a result, the canonical dimensions of the fields, and of the coupling constants, have to depend on D:

$$\begin{aligned}
\left[\int d^D x\, \mathcal{L}(x)\right] \;&=\; 0 \Rightarrow [\mathcal{L}] = D = 4-2\epsilon\;,\\
[\partial_\mu\phi\, \partial^\mu\phi] = D \;&\Rightarrow\; [\phi] = 1-\epsilon\;,\\
[\bar{\psi}\partial\!\!\!/\psi] = D \;&\Rightarrow\; [\psi] = 3/2-\epsilon\;,\\
[\bar{\psi}A\!\!\!/\psi g] = D \;&\Rightarrow\; [g] = \epsilon\;.
\end{aligned}$$

The gauge coupling constant acquires dimensions! This is a prelude to the non-trivial behaviour of the renormalized coupling constant as a function of the energy scale (“running"). But before we come to this, let us go back to the calculation of the counter-terms and the construction of the renormalized Lagrangian.

Replace the bare fields and couplings with renormalized ones:[3]

$$\begin{aligned}
\psi_{\text{bare}} &= Z_2^{1/2}\, \psi_{\rm R}\;,\\
A^\mu_{\text{bare}} &= Z_3^{1/2}\, A^\mu_{\rm R}\;,\\
g_{\text{bare}} &= Z_g\, \mu^\epsilon g_{\rm R}\;.
\end{aligned}$$

[3] For the sake of simplicity, here and in the following we shall assume the quarks to be massless. The inclusion of the mass terms does not add any interesting new feature in what follows.

We explicitly extracted the dimensions out of g_{bare}, introducing the dimensional parameter μ (renormalization scale). In this way the renormalized coupling g_{R} is dimensionless (as it should be once we go back to four dimensions).

The Lagrangian, written in terms of renormalized quantities, becomes

$$\mathcal{L} \;=\; Z_2 \bar{\psi} i \partial\!\!\!/ \psi - \frac{1}{4} Z_3 F^a_{\mu\nu} F_a^{\mu\nu} + Z_g Z_2 Z_3^{1/2} \mu^\epsilon \, g \bar{\psi} A\!\!\!/ \psi + (\text{gauge fixing, ghosts,} \dots) \,. \tag{A.27}$$

It is customary to define

$$Z_1 \;=\; Z_g \, Z_2 \, Z_3^{1/2} \,. \tag{A.28}$$

If we set $Z_n = 1 + \delta_n$, we then obtain

$$\begin{aligned} \mathcal{L} \;&=\; \bar{\psi} \, i \, \partial\!\!\!/ \psi - \frac{1}{4} F^a_{\mu\nu} F^{\mu\nu a} + \mu^\epsilon \, g \bar{\psi} A\!\!\!/ \psi + \left[\text{ghosts, gauge mixing}\right] \\ &+\; \delta_2 \, \bar{\psi} \, i \partial\!\!\!/ \psi - \frac{1}{4} \delta_3 F^a_{\mu\nu} F^{\mu\nu a} + \delta_1 \mu^\epsilon g \bar{\psi} A\!\!\!/ \psi \,. \end{aligned} \tag{A.29}$$

The counter-terms δ_i are fixed by requiring the one-loop Green functions to be finite. The explicit evaluation, which you can find carried out in detail, for example, in Refs. [3, 7], gives

$$\text{quark self-energy} \;\;\Rightarrow\;\; \delta_2 = -C_{\text{F}} \left(\frac{\alpha_s}{4\pi} \frac{1}{\epsilon} \right) \,, \tag{A.30}$$

$$\text{gluon self-energy} \;\;\Rightarrow\;\; \delta_3 = \left(\frac{5}{3} C_{\text{A}} - \frac{4}{3} n_{\text{F}} T_{\text{F}} \right) \frac{\alpha_s}{4\pi} \frac{1}{\epsilon} \,, \tag{A.31}$$

$$q\bar{q}g \text{ vertex corrections} \;\;\Rightarrow\;\; \delta_1 = -(C_{\text{A}} + C_{\text{F}}) \frac{\alpha_s}{4\pi} \frac{1}{\epsilon} \,. \tag{A.32}$$

As usual we introduced the notation $\alpha_s = g^2/4\pi$. The strong-coupling renormalization constant Z_g can be obtained using these results and Eq. (A.28):

$$Z_g \;=\; \frac{Z_1}{Z_2 Z_3^{1/2}} = 1 + \delta_1 - \delta_2 - \frac{1}{2}\delta_3 \;=\; 1 + \frac{\alpha_s}{4\pi} \frac{1}{\epsilon} \left[-\frac{11}{6} C_{\text{A}} + \frac{2}{3} \, n_{\text{F}} \, T_{\text{F}} \right] \overset{\text{def}}{\equiv} 1 - \frac{1}{\epsilon} \left(\frac{b_0}{2} \right) \alpha_s \,. \tag{A.33}$$

Note the cancellation of the terms proportional to C_{F}, between the quark self-energy (Z_2) and the Abelian part of the vertex correction (Z_1). This is the same as in the case of the QCD non-renormalization of the electric coupling, discussed at the beginning of this appendix. The non-Abelian part of the vertex correction contributes viceversa to the QCD coupling renormalization. This is a consequence of gauge invariance. The separation of the non-Abelian contributions to the self-energy and to the vertex is not gauge-invariant, only their sum is. Note also that the consistency of the renormalization procedure requires that the renormalized strong coupling g defining the strength of the interaction of quarks and gluons should be the same as that defining the interaction of gluons among themselves. If this did not happen, the gauge invariance of the $q\bar{q} \to gg$ process so painfully achieved in Section 2 by fixing the coefficient of the three-gluon coupling would no longer hold at one-loop! Once again, this additional constraint can be shown to hold through an explicit calculation.

A.4 Running of α_s

The running of α_s is a consequence of the renormalization-scale independence of the renormalization process. The bare coupling g_{bare} knows nothing about our choice of μ. The parameter μ is an artifact of the regularization prescription, introduced to define the dimensionful coupling in D dimensions, and should not enter in measurable quantities. As a result

$$\frac{d g_{\text{bare}}}{d\mu} = 0 \,. \tag{A.34}$$

Using the definition of g, $g_{\text{bare}} = \mu^\epsilon Z_g \; g$, we then get

$$\epsilon\mu^{2\epsilon} \, Z_g^2 \, \alpha_s + \mu^{2\epsilon} \, \alpha_s 2Z_g \, \frac{dZ_g}{dt} + \mu^{2\epsilon} \, Z_g^2 \, \frac{d\alpha_s}{dt} = 0 \tag{A.35}$$

where

$$\frac{d}{dt} = \mu^2 \, \frac{d}{d\mu^2} = \frac{d}{d\log\mu^2} \; . \tag{A.36}$$

Z_g depends upon μ only via the presence of α_s. If we define

$$\beta(\alpha_s) = \frac{d\alpha_s}{dt} \; , \tag{A.37}$$

we then get

$$\beta(\alpha_s) + 2\frac{\alpha_s}{Z_g} \, \frac{dZ_g}{d\alpha_s} \, \beta(\alpha_s) = -\epsilon\alpha_s \; . \tag{A.38}$$

Using Eq. (A.33) and expanding in powers of α_s, we get

$$\beta(\alpha_s) \;=\; \frac{-\epsilon\alpha_s}{1+2\frac{\alpha_s}{Z_g}\frac{dZ_g}{d\alpha_s}} \;=\; \frac{-\epsilon\alpha_s}{1-\frac{b_0\alpha_s}{\epsilon}} \;=\; -b_0\alpha_s^2 + O(\alpha_s^2,\epsilon) \tag{A.39}$$

and finally

$$\beta(\alpha_s) \;=\; -b_0\alpha_s^2 \;\text{ with }\; b_0 \;=\; \frac{1}{2\pi}\left(\frac{11}{6}C_{\text{A}} - \frac{2}{3}\,n_{\text{F}}\,T_{\text{F}}\right) \overset{N=3}{=} \frac{1}{12\pi}\,(33-2n_f) \; . \tag{A.40}$$

We can now solve Eq. (A.37), assuming $b_0 > 0$ (which is true provided the number of quark flavours is less than 16) and get the famous *running* of α_s:

$$\alpha_s(\mu^2) = \frac{1}{b_0\log(\mu^2/\Lambda^2)} \; . \tag{A.41}$$

The parameter Λ describes the boundary condition of the first-order differential equation defining the running of α_s, and corresponds to the scale at which the coupling becomes infinity.

A.5 Renormalization group invariance

The fact that the coupling constant α_s depends on the unphysical renormalization scale μ should not be a source of worry. This is because the coupling constant itself is not an observable. What we observe are decay rates, spectra, or cross-sections. These are given by the product of the coupling constant and some matrix element, which in general will acquire a non-trivial renormalization-scale dependence through the renormalization procedure. We therefore just need to check that the scale dependence of the coupling constant and of the matrix elements cancel each other, leaving results which do not depend on μ.

Consider now a physical observable, for example the ratio $R = \sigma(e^+e^- \to \text{hadrons})/\sigma(e^+e^- \to \mu^+\mu^-)$. R can be calculated in perturbation theory within QCD, giving rise to an expansion in the renormalized coupling $\alpha_s(\mu)$:

$$R[\alpha_s, s/\mu^2] \;=\; 1 + \alpha_s\, f_1(t) + \alpha_s^2\, f_2(t) + \ldots \;=\; \sum_{n=0}^{\infty} \alpha_s^n\, f_{(n)}(t) \; , \tag{A.42}$$

where $t = s/\mu^2$ (and we omitted a trivial overall factor $3\sum_f Q_f^2$). R depends on μ explicitly via the functions $f_{(n)}(t)$ and implicitly through α_s. Since R is an observable, it should be independent of μ, and the functions $f_{(n)}(t)$ cannot be totally arbitrary. In particular, one should have

$$\mu^2\frac{dR}{d\mu^2} = 0 = \left[\mu^2\frac{\partial}{\partial\mu^2} + \beta(\alpha_s)\frac{\partial}{\partial\alpha_s}\right] R[\alpha_s, s/\mu^2] = 0 \; . \tag{A.43}$$

Before we give the general, formal solution to this differential equation, it is instructive to work out directly its form within perturbation theory.

$$\mu^2 \frac{dR}{d\mu^2} \;=\; 0 \;=\; \beta(\alpha_s)\, f_1(t) + \alpha_s\, \mu^2 \frac{df_1}{d\mu^2} + 2\alpha_s\, \beta(\alpha_s)\, f_2(t) + \alpha_s^2\, \mu^2 \frac{df_2}{d\mu^2} + \ldots \tag{A.44}$$

At order α_s (remember that β is of order α_s^2) we get

$$\frac{df_1}{d\mu^2} = 0 \;\Rightarrow\; f_1 \;=\; \text{constant} \;\equiv\; a_1\ . \tag{A.45}$$

This is by itself a non-trivial result! It says that the evaluation of R at one-loop is finite, all UV infinities must cancel without charge renormalization. If they did not cancel, f_1 would depend explicitly on μ. As we saw at the beginning, this is a consequence of the non-renormalization of the electric charge.

At order α_s^2 we have

$$\beta(\alpha_s) f_1(t) + \alpha_s^2\, \frac{df_2}{d\log\mu^2} \;=\; 0 \;\Rightarrow\; f_2 = b_0\, a_1\, \log\frac{\mu^2}{s} + a_2 \text{ (integration constant)}\ . \tag{A.46}$$

So up to order α_s^2 we have

$$R = 1 + \underbrace{a_1\ \alpha_s}_{\text{one-loop}} + \underbrace{a_1\, b_0\, \alpha_s^2\ \log\ \mu^2/s + a_2\alpha_s^2}_{\text{two-loops}} + \ldots\ . \tag{A.47}$$

Note that the requirement of renormalization group invariance allows us to know the coefficient of the logarithmic term at two loops without having to carry out the explicit two-loop calculation! It is also important to notice that in the limit of high energy, $s \to \infty$, the logarithmic term of the two-loop contribution becomes very large, and this piece becomes numerically of order α_s as soon as $\log s/\mu^2 \gtrsim 1/b_0\, \alpha_s$. It is easy to check that renormalization scale invariance requires the presence of such logs at all orders of perturbation theory. In particular,

$$f_{(n)}(t) \;=\; a_1 \left[b_0 \log\frac{\mu^2}{s} \right]^n + \ldots\ . \tag{A.48}$$

We can collect all these logs as follows:

$$\begin{aligned} R \;&=\; 1 + a_1\alpha_s \left[1 + \alpha_s b_0 \log\frac{\mu^2}{s} + (\alpha_s b_0 \log\frac{\mu^2}{s})^2 + \ldots \right] + a_2\alpha_s^2 + \ldots & \text{(A.49)} \\ &=\; 1 + a_1\ \frac{\alpha_s(\mu)}{1 + \alpha_s(\mu) b_0 \log\frac{s}{\mu^2}} + a_2\alpha_s^2 + \ldots \;\equiv\; 1 + a_1\alpha_s(s) + a_2\alpha_s^2 + \ldots\ . & \text{(A.50)} \end{aligned}$$

In fact,

$$\frac{\alpha_s(\mu)}{1 + \alpha_s(\mu) b_0 \log\frac{s}{\mu^2}} = \frac{1}{b_0 \log\frac{\mu^2}{\Lambda^2} + b_0 \log\frac{s}{\mu^2}} = \frac{1}{b_0 \log\frac{s}{\Lambda^2}} \equiv \alpha_s(s)\ . \tag{A.51}$$

Renormalization group invariance constrains the form of higher-order corrections. All of the higher-order logarithmic terms are determined in terms of lower-order finite coefficients. They can be resummed by simply setting the scale of α_s to s. You can check by yourself that this will work also for the higher-order terms, such as those proportional to a_2. So the final result has the form

$$R = 1 + a_1\alpha_s(s) + a_2\alpha_s^2(s) + a_3\alpha_s^3(s) + \ldots\ . \tag{A.52}$$

Of course $a_1, a_2, \ldots$ have to be determined by an explicit calculation. However, the truncation of the series at order n has now an accuracy which is truly of order α_s^{n+1}, contrary to before when higher-order terms were as large as lower-order ones. The explicit calculation has been carried out up to the a_3 coefficient. In particular,

$$a_1 = \frac{3}{4}\ \frac{C_{\rm F}}{\pi} \equiv \frac{1}{\pi}\ . \tag{A.53}$$

The formal proof of the previous equation can be obtained by showing that the general form of the equation

$$\left[\mu^2 \ \frac{\partial}{\partial\mu^2} + \beta(\alpha_s)\right] \ R\,(\alpha_s, \frac{s}{\mu^2}) = 0 \tag{A.54}$$

is given by

$$\begin{cases} R(\alpha_s(s), 1) \ , \ \text{with} \\ \dfrac{d\alpha_s}{d\log\frac{s}{\mu^2}} = \beta(\alpha_s) \end{cases} . \tag{A.55}$$

B Formal derivation of the evolution equations

Assuming the parton picture outlined above, we can describe the cross-section for the interaction of the virtual photon with the proton as follows:

$$\sigma_0 = \int_0^1 dx \sum_i e_i^2 \, f_i(x) \, \hat\sigma_0(\gamma^* q_i \to q_i', \, x) \tag{B.1}$$

where the 0 subscript anticipates that this description represents a leading order approximation. In the above equation, $f_i(x)$ represents the density of quarks of flavour i carrying a fraction x of the proton momentum. The hatted cross-section represents the interaction between the photon and a free (massless) quark:

$$\begin{aligned} \hat\sigma_0(\gamma^* q_i \to q_i') &= \frac{1}{\text{flux}} \overline{\sum} \, |M_0(\gamma^* q \to q')|^2 \frac{d^3p'}{(2\pi)^3 2p_0'} \ (2\pi)^4 \delta^4(p' - q - p) \\ &= \frac{1}{\text{flux}} \overline{\sum} |M_0|^2 2\pi\delta(p'^2) \ . \end{aligned} \tag{B.2}$$

Using $p' = xP + q$, where P is the proton momentum, we get

$$(p')^2 = 2xP \cdot q + q^2 \equiv 2xP \cdot q - Q^2 \ , \tag{B.3}$$

$$\hat\sigma_0(\gamma^* q \to q') = \frac{2\pi}{\text{flux}} \overline{\sum} |M_0|^2 \, \frac{1}{2P \cdot q} \, \delta(x - x_{bj}) \ , \tag{B.4}$$

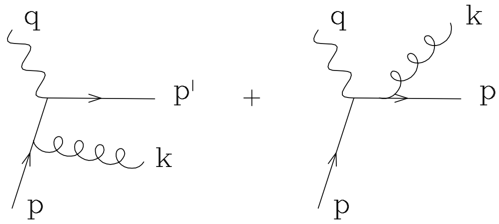
where $x_{bj} = \frac{Q^2}{2P \cdot q}$ is the so-called Bjorken-x variable. Finally,

$$\sigma_0 = \frac{2\pi}{\text{flux}} \, \frac{\overline{\sum}|M_0|^2}{Q^2} \sum_i x_{bj} \, f_i(x_{bj}) \, e_i^2 \ \equiv \ \frac{2\pi}{\text{flux}} \, \frac{\overline{\sum}|M_0|^2}{Q^2} \, F_2(x_{bj}) \ . \tag{B.5}$$

The measurement of the inclusive ep cross-section as a function of Q^2 and $P \cdot q$ [$= m_p(E' - E)$ in the proton rest frame, with E' = energy of final-state lepton and E = energy of initial-state lepton] probes the quark momentum distribution inside the proton.

B.1 Parton evolution

Let us now study the QCD corrections to the LO parton-model description of DIS. This study will exhibit many important aspects of QCD (structure of collinear singularities, renormalization-group invariance) and will take us to an important element of the DIS phenomenology, namely scaling violations. We start from real-emission corrections to the Born level process:



(B.6)

The first diagram is proportional to $1/(p-k)^2 = 1/2(pk)$, which diverges when k is emitted parallel to p:

$$p \cdot k = p^0 k^0 \, (1 - \cos\theta) \overset{\cos\theta \to 1}{\longrightarrow} 0 \, . \tag{B.7}$$

The second diagram is also divergent, if k is emitted parallel to p'. This second divergence turns out to be harmless, since we are summing over all possible final states. Whether the final-state quark keeps all of its energy, or whether it decides to share it with a gluon emitted collinearly, an inclusive final-state measurement will not care. The collinear divergence can then be cancelled by a similar divergence appearing in the final-state quark self-energy corrections.

The first divergence is more serious, since from the point of view of the incoming photon (which only sees the quark, not the gluon) it *does* make a difference whether the momentum is all carried by the quark or is shared between the quark and the gluon. This means that no cancellation between collinear singularities in the real emission and virtual emission is possible. So let us go ahead, calculate explicitly the contribution of these diagrams, and learn how to deal with their singularities.

First of all, note that while the second diagram is not singular in the region $k{\cdot}p \to 0$, its interference with the first one is. It is possible, however, to select a gauge for which the interference of the two diagrams is finite in this limit. It can be shown that the right choice is

$$\sum \epsilon_\mu \epsilon^*_\nu(k) = -g_{\mu\nu} + \frac{k^\mu p'^\nu + k^\nu p'^\mu}{k \cdot p'} \, . \tag{B.8}$$

Note that in this gauge not only $k \cdot \epsilon(k) = 0$, but also $p' \cdot \epsilon(k) = 0$. The key to getting to the end of a QCD calculation in a finite amount of time is choosing a proper gauge (which we just did) and the proper parametrization of the momenta involved. In our case, since we are interested in isolating the region where k becomes parallel with p, it is useful to set

$$k_\mu = (1-z)p_\mu + \beta p'_\mu + (k_\perp)_\mu \, , \tag{B.9}$$

with $k_\perp \cdot p = k_\perp \cdot p' = 0$. β is obtained by imposing

$$k^2 = 0 = 2\beta(1-z) p \cdot p' + k_\perp^2 \, . \tag{B.10}$$

Defining $k_\perp^2 = -k_t^2$, we then get

$$\beta = \frac{k_t^2}{2(pp')\,(1-z)} \, , \tag{B.11}$$

$$k_\mu = (1-x)p_\mu + \frac{k_t^2}{2(1-x)p \cdot p'}\, p'_\mu + (k_\perp)_\mu \, . \tag{B.12}$$

$(k_\perp)_\mu$ is therefore the gluon momentum vector transverse to the incoming quark, in a frame where γ^* and q are aligned. k_t is the value of this transverse momentum. We also get

$$k \cdot p \; = \; \beta \, p \cdot p' = \frac{k_t^2}{2(1-z)} \quad \text{and} \quad k \cdot p' \; = \; (1-z) p \cdot p' \, . \tag{B.13}$$

As a result $(p-k)^2 = -k_t^2/(1-z)$. The amplitude for the only diagram carrying the initial-state singularity is

$$M_g \; = \; ig\lambda^a_{ij} \, \bar{u}(p')\Gamma \, \frac{\hat{p} - \hat{k}}{(p-k)^2} \, \hat{\epsilon}(k) u(p) \tag{B.14}$$

(where we introduced the notation $\hat{a} \equiv \not{a} \equiv a_\mu \gamma^\mu$). We indicated by Γ the interaction vertex with the external current q. It is important to keep Γ arbitrary, because we would like to get results which do not depend on the details of the interaction with the external probe. It is important that the singular part of the QCD correction, and therefore its renormalization, be process independent. Only in this way can we hope to achieve a true universality of the parton densities! So we will keep Γ generic, and make sure that our algebra does not depend on its form, at least in the $p \cdot k \to 0$ limit. Squaring the most singular part of the amplitude, and summing over colours and spins, we get

$$\sum_{\substack{\text{g polariz.}\\ \text{and colours}}} |M_g|^2 \;=\; g^2 \overbrace{\sum_a \text{tr}\,(\lambda^a\lambda^a)}^{N\times C_{\rm F}} \times \frac{1}{t^2} \times \sum_\epsilon \text{tr}\,[\hat{p}'\,\Gamma(\hat{p}-\hat{k})\,\hat{\epsilon}\,p\,\hat{\epsilon}^*\,(\hat{p}-\hat{k})\,\Gamma^+] \tag{B.15}$$

with $t = (p-k)^2 = -k_t^2/(1-z)$. Let us look first at

$$\sum_\epsilon \hat{\epsilon}\,\hat{p}\,\hat{\epsilon}^* \;=\; \sum \epsilon_\mu\,\epsilon^*_\nu\;\gamma^\mu\hat{p}\gamma^\nu \;=\; -\gamma^\mu\hat{p}\gamma^\mu + \frac{1}{k\cdot p'}\,(\hat{p}'\hat{p}\hat{k} + \hat{k}\hat{p}\hat{p}') \;=\; \frac{2}{1-z}\,(\hat{k}+\beta\hat{p}') \tag{B.16}$$

(we used $\hat{a}\hat{b}\hat{c} + \hat{c}\hat{b}\hat{a} = 2(a\cdot b)\,\hat{c} - 2(a\cdot c)\,\hat{b} + 2(b\cdot c)\,\hat{a}$ and some of the kinematical relations from the previous page). Then take

$$(\hat{p}-\hat{k})\,(\hat{k}+\beta\hat{p}')\,(\hat{p}-\hat{k}) = (\hat{p}-\hat{k})\,\hat{k}\,(\hat{p}-\hat{k}) + \beta(\hat{p}-\hat{k})\,\hat{p}'\,(\hat{p}-\hat{k})\;. \tag{B.17}$$

In the second term, proportional to β, we can approximate $\hat{k} = (1-z)\hat{p}$. This is because the other pieces $(\beta\hat{p}' + \hat{k}_\perp)$ multiplied by β would cancel entirely the $\frac{1}{t^2}$ singularity, and would only contribute a non-singular term, which we are currently neglecting. So Eq. (B.17) becomes

$$\hat{p}\hat{k}\hat{p} + \beta z^2\hat{p}\hat{p}'\hat{p} = 2(p\cdot k)\hat{p} + \beta z^2 2(p\cdot p')\hat{p} = 2(p\cdot k)\,(1+z^2)\hat{p} \tag{B.18}$$

and

$$\sum |M_g|^2 = 2g^2\,C_{\rm F}\,\frac{(1-z)}{k_t^2}\,\left(\frac{1+z^2}{1-z}\right)\;N\,\text{tr}[\hat{p}'\Gamma\hat{p}\Gamma^+]\;. \tag{B.19}$$

The last factor with the trace corresponds to the Born amplitude squared. So the one-gluon emission process factorizes in the collinear limit into the Born process times a factor which is independent of the beam's nature! If we add the gluon phase-space

$$[dk] \equiv \frac{d^3k}{(2\pi)^3 2k^0} = \frac{dk_\parallel}{k^0}\,\frac{d\phi}{2\pi}\,\frac{1}{8\pi^2}\,\frac{dk_\perp^2}{2} = \frac{dz}{(1-z)}\,\frac{1}{16\pi^2}\,dk_\perp^2\;, \tag{B.20}$$

we get

$$\overline{\sum}|M_g|^2\,[dk] = \frac{dk_\perp^2}{k_\perp^2}\,dz\,\left(\frac{\alpha_s}{2\pi}\right)\,P_{qq}(z)\,\overline{\sum}|M_0|^2 \tag{B.21}$$

where

$$P_{qq}(z) = C_{\rm F}\frac{1+z^2}{1-z} \tag{B.22}$$

is the so-called Altarelli–Parisi splitting function for the $q \to q$ transition (z is the momentum fraction of the original quark taken away by the quark after gluon emission). We are now ready to calculate the corrections to the parton-model cross-section:

$$\sigma_g \;=\; \int dx\,f(x)\,\frac{1}{\text{flux}}\,\int dz\,\frac{dk_\perp^2}{k_\perp^2}\,\left(\frac{\alpha_s}{2\pi}\right)\,P_{qq}(z)\,\overline{\sum}|M_0|^2\,2\pi\delta(p'^2)\;. \tag{B.23}$$

Using $(p')^2 = (p-k+q)^2 \sim (zp+q)^2 = (xzP+q)^2$ and

$$\delta(p'^2) \;=\; \frac{1}{2P\cdot q}\,\frac{1}{z}\,\delta(x - \frac{x_{bj}}{z}) \;=\; \frac{x_{bj}}{z}\,\delta(x - \frac{x_{bj}}{z})\;, \tag{B.24}$$

we finally obtain

$$\sigma_g = \frac{2\pi}{\text{flux}} \left(\frac{\overline{\sum}|M_0|^2}{Q^2} \right) \sum_i e_i^2 \, x_{bj} \, \frac{\alpha_s}{2\pi} \int \frac{dk_\perp^2}{k_\perp^2} \int \frac{dz}{z} \, P_{qq}(z) \, f_i \left(\frac{x_{bj}}{z} \right) . \tag{B.25}$$

We then find that the inclusion of the $\mathcal{O}(\alpha_s)$ correction is equivalent to a contribution to the parton density:

$$f_i(x) \to f_i(x) + \frac{\alpha_s}{2\pi} \int \frac{dk_\perp^2}{k_\perp^2} \int_x^1 \frac{dz}{z} \, P_{qq}(z) \, f_i \left(\frac{x}{z} \right) . \tag{B.26}$$

Note the presence of the integral $\int dk_\perp^2/k_\perp^2$. The upper limit of integration is proportional to Q^2. The lower limit is 0. Had we included a quark mass, the propagator would have behaved like $1/(k_\perp^2 + m^2)$. But the quark is bound inside the hadron, so we do not quite know what m should be. Let us then assume that we cut off the integral at a $k_\perp$ value equal to some scale μ_0, and see what happens. The effective parton density becomes

$$f(x, Q^2) = f(x) + \log \left(\frac{Q^2}{\mu_0^2} \right) \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \, P_{qq}(z) \, f \left(\frac{x}{z} \right) . \tag{B.27}$$

The dependence on the scale μ_0, which is a non-perturbative scale, can be removed by defining $f(x, Q^2)$ in terms of the parton density f *measured* at a large, perturbative scale μ^2:

$$f(x, \mu^2) = f(x) + \log \left(\frac{\mu^2}{\mu_0^2} \right) \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \, P_{qq}(z) \, f \left(\frac{x}{z} \right) . \tag{B.28}$$

We can then perform a subtraction, and write

$$f(x, Q^2) = f(x, \mu^2) + \log \left(\frac{Q^2}{\mu^2} \right) \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \, P_{qq}(z) \, f \left(\frac{x}{z} \right) . \tag{B.29}$$

The scale μ plays here a similar role to the renormalization scale introduced in the Appendix A. Its choice is arbitrary, and $f(x, Q^2)$ should not depend on it. Requiring this independence, we get the following ‘renormalization-group (RG) invariance’ condition:

$$\frac{df(x, Q^2)}{d \ln \mu^2} = \mu^2 \, \frac{df(x, \mu^2)}{d\mu^2} - \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \, P_{qq}(z) \, f \left(\frac{x}{z} \right) \equiv 0 \tag{B.30}$$

and then

$$\mu^2 \frac{df(x, \mu^2)}{d\mu^2} \;=\; \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \, P_{qq}(z) \, f \left(\frac{x}{z}, \mu^2 \right) . \tag{B.31}$$

This equation is usually called the DGLAP (Dokshitzer–Gribov–Lipatov–Altarelli–Parisi) equation. As in the case of the resummation of leading logarithms in $R_{e^+e^-}$ induced by the RG invariance constraints, the DGLAP equation—which is the result of RG invariance—resums a full tower of leading logarithms of Q^2.

Proof: Let us define $t = \log \frac{Q^2}{\mu^2}$. We can then expand $f(x, t)$ in powers of t:

$$f(x, t) \;=\; f(x, 0) + t \frac{df}{dt}(x, 0) + \frac{t^2}{2!} \, \frac{d^2 f}{dt^2}(x, 0) + \ldots \, . \tag{B.32}$$

The first derivative is given by the DGLAP equation itself. Higher derivatives can be obtained by differentiating it:

$$f''(x, t) \;=\; \frac{\alpha_s}{2\pi} \int \frac{dz}{z} P_{qq}(z) \frac{df}{dt}(\frac{x}{z}, t)$$

$$
\begin{aligned}
&= \frac{\alpha_s}{2\pi}\int_x^1 \frac{dz}{z} P_{qq}(z) \frac{\alpha_s}{2\pi}\int_{\frac{x}{z}}^1 \frac{dz'}{z'} P_{qq}(z) f(\frac{x}{zz'}, t) \\
&\vdots \\
f^{(h)}(x,t) &= \frac{\alpha_s}{2\pi}\int_x^1 \ldots\ldots \frac{\alpha_s}{2\pi}\int_{x/zz'\ldots z^{(n-1)}}^1 \frac{dz^{(n)}}{z^{(n)}} P_{qq}(z^{(n)}) f(\frac{x}{zz'\ldots}, t) \, . \qquad \text{(B.33)}
\end{aligned}
$$

The n-th term in this expansion, proportional to $(\alpha_s\, t)^n$, corresponds to the emission of n gluons (it is just the n-fold iteration of what we did studying the one-gluon emission case).

With similar calculations one can include the effect of the other $\mathcal{O}(\alpha_s)$ correction, originating from the splitting into a $q\bar{q}$ pair of a gluon contained in the proton. With the addition of this term, the evolution equation for the density of the ith quark flavour becomes

$$
\frac{df_q(x,t)}{dt} = \frac{\alpha_s}{2\pi}\int_x^1 \frac{dz}{z}\left[P_{qq}(z)\, f_i(\frac{x}{z},t) \,+\, P_{qg}(z) f_g(\frac{x}{z},t)\right] \, , \quad \text{with } P_{qg} = \frac{1}{2}\left[z^2 + (1-z)^2\right] \, . \qquad \text{(B.34)}
$$

In the case of interactions with a coloured probe (say a gluon) we meet the following corrections, which affect the evolution of the gluon density $f_g(x)$:

$$
\frac{df_g(x,t)}{dt} = \frac{\alpha_s}{2\pi}\int_x^1 \frac{dz}{z}\left[P_{gq}(z)\sum_{i=q,\bar{q}} f_i\left(\frac{x}{z},t\right) + P_{gg}(z) f_g\left(\frac{x}{z},t\right)\right] \qquad \text{(B.35)}
$$

with

$$
P_{gq}(z) = P_{qq}(1-z) = C_\text{F}\frac{1+(1-z)^2}{z} \quad \text{and} \quad P_{gg}(z) = 2C_\text{A}\left[\frac{1-z}{z} + \frac{z}{1-z} + z(1-z)\right] \, . \qquad \text{(B.36)}
$$

Defining the moments of an arbitrary function $g(x)$ as follows,

$$
g_n = \int_0^1 \frac{dx}{x}\, x^n\, g(x) \, ,
$$

it is easy to prove that the evolution equations turn into ordinary linear differential equations:

$$
\begin{aligned}
\frac{df_i^{(n)}}{dt} &= \frac{\alpha_s}{2\pi}[P_{qq}^{(n)} f_i^{(n)} + P_{qg}^{(n)} f_g^{(n)}] \, , && \text{(B.37)} \\
\frac{df_g^{(n)}}{dt} &= \frac{\alpha_s}{2\pi}[P_{gg}^{(n)} f_g + P_{gq}^{(n)} f_i^{(n)}] \, . && \text{(B.38)}
\end{aligned}
$$

B.2 Properties of the evolution equations

We now study some general properties of these equations. It is convenient to introduce the concepts of *valence* $[V(x,t)]$ and *singlet* $[\Sigma(x,t)]$ densities:

$$
\begin{aligned}
V(x) &= \sum_i f_i(x) - \sum_{\bar{\imath}} f_{\bar{\imath}}(x) \, , && \text{(B.39)} \\
\Sigma(x) &= \sum_i f_i(x) + \sum_{\bar{\imath}} f_{\bar{\imath}}(x) \, , && \text{(B.40)}
\end{aligned}
$$

where the index $\bar{\imath}$ refers to the antiquark flavours. The evolution equations then become

$$
\frac{dV^{(n)}}{dt} = \frac{\alpha_s}{2\pi} P_{qq}^{(n)}\, V^{(n)} \, , \qquad \text{(B.41)}
$$

$$\frac{d\Sigma^{(n)}}{dt} = \frac{\alpha_s}{2\pi}\left[P_{qq}^{(n)}\,\Sigma^{(n)} + 2n_f\,P_{qg}^{(n)}\,f_g^{(n)}\right], \quad \text{(B.42)}$$

$$\frac{df_g^{(n)}}{dt} = \frac{\alpha_s}{2\pi}\left[P_{gq}^{(n)}\,\Sigma^{(n)} + P_{gg}^{(n)}\,f_g^{(n)}\right]. \quad \text{(B.43)}$$

Note that the equation for the valence density decouples from the evolution of the gluon and singlet densities, which are coupled among themselves. This is physically very reasonable, since in perturbation theory the contribution to the quark and the antiquark densities coming from the evolution of gluons (via their splitting into $q\bar{q}$ pairs) is the same, and will cancel out in the definition of the valence. The valence therefore only evolves because of gluon emission. On the contrary, gluons and $q\bar{q}$ pairs in the proton *sea* evolve into one another.

The first moment of $V(x)$, $V^{(1)} = \int_0^1 dx\,V(x)$, counts the number of valence quarks. We therefore expect it to be independent of Q^2:

$$\frac{dV^{(1)}}{dt} \equiv 0 = \frac{\alpha_s}{2\pi}\,P_{qq}^{(1)}\,V^{(1)} = 0\,. \quad \text{(B.44)}$$

Since $V^{(1)}$ itself is different from 0, we obtain a constraint on the first moment of the splitting function: $P_{qq}^{(1)} = 0$. This constraint is satisfied by including the effect of the virtual corrections, which generate a contribution to $P_{qq}(z)$ proportional to $\delta(1-z)$. This correction is incorporated in $P_{qq}(z)$ via the redefinition:

$$P_{qq}(z) \to \left(\frac{1+z^2}{1-z}\right)_+ \equiv \frac{1+z^2}{1-z} - \delta(1-z)\int_0^1 dy\,\left(\frac{1+y^2}{1-y}\right) \quad \text{(B.45)}$$

where the + sign turns $P_{qq}(z)$ into a distribution. In this way, $\int_0^1 dz\,P_{qq}(z) = 0$ and the valence sum-rule is obeyed at all Q^2.

Another sum-rule that does not depend on Q^2 is the momentum sum-rule, which imposes the constraint that all of the momentum of the proton is carried by its constituents (valence plus sea plus gluons):

$$\int_0^1 dx\,x\left[\sum_{i,\bar{i}} f_i(x) + f_g(x)\right] \equiv \Sigma^{(2)} + f_g^{(2)} = 1\,. \quad \text{(B.46)}$$

Once more this relation should hold for all Q^2 values, and this can be proved by using the evolution equations that this implies:

$$P_{qq}^{(2)} + P_{gq}^{(2)} = 0\,, \quad \text{(B.47)}$$

$$P_{gg}^{(2)} + 2n_f\,P_{qg}^{(2)} = 0\,. \quad \text{(B.48)}$$

You can check using the definition of second moment, and the explicit expressions of the P_{qq} and P_{gq} splitting functions, that the first condition is automatically satisfied. The second condition is satisfied by including the virtual effects in the gluon propagator, which contribute a term proportional to $\delta(1-z)$. It is a simple exercise to verify that the final form of the $P_{gg}(z)$ splitting function, satisfying Eq. (B.48), is

$$P_{gg} \to 2C_{\mathrm{A}}\left\{\frac{x}{(1-x)_+} + \frac{1-x}{x} + x(1-x)\right\} + \delta(1-x)\left[\frac{11C_{\mathrm{A}} - 2n_f}{6}\right]. \quad \text{(B.49)}$$

B.3 Solution of the evolution equations

The evolution equations formulated in the previous section can be solved analytically in moment space. The boundary conditions are given by the moments of the parton densities at a given scale μ, where in principle they can be obtained from a direct measurement. The solution at different values of the scale Q can then be obtained by inverting numerically the expression for the moments back to x space. The resulting evolved densities can then be used to calculate cross-sections for an arbitrary process involving hadrons, at an arbitrary scale Q. We shall limit ourselves here to studying some properties of the analytic solutions, and will present and comment on some plots obtained from numerical studies available in the literature.

As an exercise, you can show that the solution of the evolution equation for the valence density is the following:

$$V^{(n)}(Q^2) \;=\; V^{(n)}(\mu^2)\left[\frac{\log Q^2/\Lambda^2}{\log \mu^2/\Lambda^2}\right]^{P_{qq}^{(n)}/2\pi b_0} \;=\; V^{(n)}(\mu^2)\left[\frac{\alpha_s(\mu^2)}{\alpha_s(Q^2)}\right]^{P_{qq}^{(n)}/2\pi b_0} \tag{B.50}$$

where the running of $\alpha_s(\mu^2)$ has to be taken into account to get the right result. Since all moments $P^{(n)}$ are negative, the evolution to larger values of Q makes the valence distribution softer and softer. This is physically reasonable, since the only thing that the valence quarks can do is to lose energy because of gluon emission.

The solutions for the gluon and singlet distributions f_g and Σ can be obtained by diagonalizing the 2×2 system in Eqs. (B.42) and (B.43). We study the case of the second moments, which correspond to the momentum fractions carried by quarks and gluons separately. In the asymptotic limit $\Sigma^{(2)}$ goes to a constant, and $\frac{d\Sigma^{(2)}}{dt} = 0$. Then, using the momentum sum-rule,

$$P_{qq}^{(2)}\,\Sigma^{(2)} + 2n_f\,P_{qg}^{(2)} f_g^{(2)} \;=\; 0\,, \tag{B.51}$$

$$\Sigma^{(2)} + f_g^{(2)} \;=\; 1\,. \tag{B.52}$$

The solution of this system is

$$\Sigma^{(2)} = \frac{1}{1+\frac{4C_\mathrm{F}}{n_f}} \qquad (= 15/31 \text{ for } n_f = 5)\,, \tag{B.53}$$

$$f_g^{(2)} = \frac{4C_\mathrm{F}}{4C_\mathrm{F}+n_f} \qquad (= 16/31 \text{ for } n_f = 5)\,. \tag{B.54}$$

As a result, the fraction of momentum carried by gluons is asymptotically approximately 50% of the total proton momentum. It is interesting to note that, experimentally, this asymptotic value is actually reached already at rather low values of Q^2. It was indeed observed already since the early days of the DIS experiments that only approximately 50% of the proton momentum was carried by charged constituents. This was one of the early pieces of evidence for the existence of gluons.

A complete solution for the evolved parton densities in x space can only be obtained from a numerical analysis. This work has been done in the past by several groups (see e.g., the discussions in Ref. [8]), and is continuously being updated by including the most up-to-date experimental results used for the determination of the input densities at a fixed scale.

C Jet rates in e^+e^- collisions

We present here explicit calculations of a few interesting jet observables in e^+e^- collisions. For simplicity, we will work with the soft-gluon approximation for the matrix elements and the phase-space. As a result, the correction to the differential $e^+e^- \to q\bar{q}$ cross-section from one-gluon emission becomes

$$d\sigma_g = \sigma_0\,\frac{2\alpha_s}{\pi}\;C_\mathrm{F}\;\frac{dk_0}{k_0}\;\frac{d\cos\theta}{1-\cos^2\theta} \quad , \quad \text{where } \sigma_0 \text{ is the Born amplitude}\,. \tag{C.1}$$

In this equation we used the fact that in the soft-g limit the q and $\bar{q}$ are back-to-back, and

$$q\cdot\bar{q} = 2q_0\bar{q}_0\,, \quad q\cdot k = q_0k_0(1-\cos\theta), \quad \bar{q}k = \bar{q}_0k_0(1+\cos\theta)\,. \tag{C.2}$$

Note the presence in $d\sigma_g$ of soft and collinear singularities. They will have to cancel in the total cross-section which, as we saw in the previous lecture, is finite. They do indeed cancel against the contribution to the total cross-section coming from the virtual correction diagram, where a gluon is exchanged between the two quarks. In the total cross-section (and for other sufficiently inclusive observables) the final states produced by the virtual diagrams and by the real emission diagrams in the soft or collinear limit are the same, and both contribute. In order for the total cross-section to be finite, the virtual contribution will need to take the following form:

$$\frac{d^2\sigma_v}{dk_0 d\cos\theta} = -\sigma_0 \frac{2\alpha_s}{\pi} C_{\rm F} \int_0^{\sqrt{s/2}} \frac{dk_0'}{k_0'} \int_{-1}^{1} \frac{d\cos\theta'}{1-\cos^2\theta'} \times \frac{1}{2}\delta(k_0)\, [\delta(1-\cos\theta)+\delta(1+\cos\theta)] \tag{C.3}$$

plus finite corrections. In this way,

$$\int_0^{\sqrt{s/2}} dk_0 \int_{-1}^{1} d\cos\theta \left[\frac{d^2\sigma_g}{dk_0 d\cos\theta} + \frac{d^2\sigma_v}{dk_0 d\cos\theta}\right] = \text{finite}\,. \tag{C.4}$$

With the form of the virtual corrections available (at least in this simplified soft-gluon-dominated approximation), we can proceed and calculate other quantities.

Jets are usually defined as clusters of particles close-by in phase-space. A typical jet definition distributes particles in sets of invariant mass smaller than a given parameter M, requiring that one particle only belongs to one jet, and that no other particles (or jets) can be added to a given jet without its mass exceeding M. In the case of a three-particle final state, such as the one we are studying, we get three-jet events if $(q+k)^2$, $(\bar{q}+k)^2$ and $(q+\bar{q})^2$ are all larger than M^2. We will have two-jet events when at least one of these quantities gets smaller than M^2. For example emission of a gluon near the direction of the quark, with $2qk = 2q^0k^0(1-\cos\theta) < M^2$, defines a two-jet event, one jet being given by the $\bar{q}$, the other by the system $q+k$.

One usually introduces the parameter $y = M^2/s$, and studies the jet multiplicity as a function of y. Let us calculate the two- and three-jet rates at order α_s. The phase-space domain for two-jet events is given by two regions. The first one is defined by $2qk = 2q_0k_0(1-\cos\theta) < ys$. This region consists of two parts:

$$\text{(I)}_a: \quad \begin{cases} k_0 < y\sqrt{s} \\ 0 < \cos\theta < 1 \end{cases} \qquad \oplus \qquad \text{(I)}_b: \quad \begin{cases} k_0 > y\sqrt{s} \\ 1-\frac{y\sqrt{s}}{k_0} < \cos\theta < 1 \end{cases} \quad . \tag{C.5}$$

(I)$_a$ corresponds to soft gluons at all angles smaller than $\pi/2$ (i.e., in the quark emisphere), and (I)$_b$ corresponds to hard gluons emitted at small angles from the quark.

The second region, (II), is analogous to (I), but the angles are now referred to the direction of the antiquark. The integrals of $d\sigma$ over (I) and (II) are of course the same. The $\mathcal{O}(\alpha_s)$ contribution to the two-jet rate is therefore given by

$$\begin{aligned}
\frac{\sigma^{(\alpha_s)}_{\text{2-jet}}}{\sigma_0} &= \frac{1}{\sigma_0}\left[2\int_{(I)_a} d\sigma_g + 2\int_{(I)_b} d\sigma_g + \int_{\rm virtual} d\sigma_v\right] \\
&= \frac{4\alpha_s C_{\rm F}}{\pi}\left[\int_0^{y\sqrt{s}} \frac{dk_0}{k_0}\int_0^1 \frac{d\cos\theta}{1-\cos^2\theta} + \int_{y\sqrt{s}}^{\sqrt{s/2}} \frac{dk_0}{k_0}\int_{1-(\frac{y\sqrt{s}}{k_0})}^{1} \frac{d\cos\theta}{1-\cos^2\theta}\right. \\
&\qquad \left. - \int_0^{y\sqrt{s}} \frac{dk_0}{k_0}\int_0^1 \frac{d\cos\theta}{1-\cos^2\theta}\right] \\
&= \frac{4\alpha_s C_{\rm F}}{\pi}\left\{-\int_{y\sqrt{s}}^{\sqrt{s/2}} \frac{dk_0}{k_0}\int_0^1 \frac{d\cos\theta}{1-\cos^2\theta} + \int_{y\sqrt{s}}^{\sqrt{s/2}} \frac{dk_0}{k_0}\int_{1-(\frac{y\sqrt{s}}{k_0})}^{1} \frac{d\cos\theta}{1-\cos^2\theta}\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{4\alpha_s C_{\rm F}}{\pi} \int_{y\sqrt{s}}^{\sqrt{s}/2} \frac{dk_0}{k_0} \int_0^{1-(\frac{y\sqrt{s}}{k_0})} \left(\frac{d\cos\theta}{1-\cos^2\theta} \right) \\
&= \frac{2\alpha_s C_{\rm F}}{\pi} \int_{y\sqrt{s}}^{\sqrt{s}/2} \frac{dk_0}{k_0} \left[(-)\log\frac{k_0}{y\sqrt{s}} + (\text{finite for } y \to 0) \right] = -\frac{\alpha_s C_{\rm F}}{\pi} \log^2 2y \; . \qquad \text{(C.6)}
\end{aligned}
$$

Including the Born contribution, which always gives rise to two and only two jets, we finally have

$$
\begin{aligned}
\sigma_{2-{\rm jet}} &= \sigma_0 \left[1 - \frac{\alpha_s C_{\rm F}}{\pi} \log^2 y + \ldots \right] , \\
\sigma_{3-{\rm jet}} &= \sigma_0 \frac{\alpha_s C_{\rm F}}{\pi} \log^2 y + \ldots \; .
\end{aligned}
$$

If $y \to 0$, $\sigma_{3-{\rm jet}}$ becomes larger than $\sigma_{2-{\rm jet}}$. If y is sufficiently small, we can even get $\sigma_{2-{\rm jet}} < 0$! This is a sign that higher-order corrections become important. In the soft-gluon limit, assuming that the emission of a second gluon will also factorize[4], we can repeat the calculation at higher orders and obtain

$$
\begin{aligned}
\sigma_{2-{\rm jet}} &\simeq \sigma_0 \left[1 - \frac{\alpha_s C_{\rm F}}{\pi} \log^2 y + \frac{1}{2!} \left(\frac{\alpha_s C_{\rm F}}{\pi} \log^2 y \right)^2 + \ldots \right] = \sigma_0 \, e^{-\frac{\alpha_s C_{\rm F}}{\pi} \log^2 y} , \\
\sigma_{3-{\rm jet}} &\sim \sigma_0 \, \frac{\alpha_s C_{\rm F}}{\pi} \log^2 y \; e^{-\frac{\alpha_s C_{\rm F}}{\pi} \log^2 y} , \\
&\vdots \\
\sigma_{(n+2)-{\rm jet}} &\sim \sigma_0 \, \frac{1}{n!} \left(\frac{\alpha_s C_{\rm F}}{\pi} \log^2 y \right)^n e^{-\frac{\alpha_s C_{\rm F}}{\pi} \log^2 y} \; . \qquad \text{(C.7)}
\end{aligned}
$$

It is immediate to recognize in this series a Poisson distribution, leading to an average number of jets given by

$$
\langle n_{\rm jet} \rangle \simeq 2 + \frac{\alpha_s C_{\rm F}}{\pi} \log^2 y \; . \qquad \text{(C.8)}
$$

The *smaller* the resolution parameter y, the *smaller* the mass of the jets, and the *larger* the importance of higher-order corrections. If we take the parameter M down to the scale of a few hundred MeV ($M \sim \Lambda_{\rm QCD}$), each particle gets identified with an independent jet. We can therefore estimate the s-dependence of the average multiplicity of particles produced:

$$
\langle n_{\rm part} \rangle \sim \frac{C_{\rm F}\alpha_s}{\pi} \log^2 \frac{s}{\Lambda^2} = \frac{C_{\rm F}}{\pi} \frac{1}{b_0 \log\frac{s}{\Lambda^2}} \log^2 \frac{s}{\Lambda^2} \simeq \frac{C_{\rm F}}{\pi b_0} \log \frac{s}{\Lambda^2} \; . \qquad \text{(C.9)}
$$

The final-state particle multiplicity grows with $\log(s)$.

In practice, things are a bit more complicated than this. Once the first gluon is emitted, additional gluons can be emitted from it as well. Therefore the final-state multiplicity will be dominated by the emission of gluons from gluons. The analysis becomes more complicated (see e.g., Refs. [6] and [8] for the details), and the final result is

$$
\langle n_{\rm part}(s) \rangle \sim \exp \sqrt{\frac{2C_{\rm A}}{\pi b} \log(\frac{s}{\Lambda^2})} \qquad \text{(C.10)}
$$

for the particle multiplicity, and

$$
\langle n_{\rm jet}(y) \rangle = 2 + 2\frac{C_{\rm F}}{C_{\rm A}} (\cosh \sqrt{\frac{\alpha_s C_{\rm A}}{2\pi} \log^2 \frac{1}{y}} - 1) \sim \frac{C_{\rm F}}{C_{\rm A}} \exp \sqrt{\frac{\alpha_s C_{\rm A}}{2\pi} \log^2 \frac{1}{y}}
$$

[4] This is not true (see later on), but let us just accept it to see how things develop.

for the average jet multiplicity.

Other interesting quantities that can be calculated using the simple formulas we developed so far are the average jet mass and the thrust. To define the jet mass we just divide the final state into two emispheres, separated by the plane orthogonal to the thrust axis. We now call jets the two sets of particles on either side of the plane. The $\langle m^2 \rangle$ of the jet is then given by

$$\langle m^2_{\rm jet} \rangle = \frac{1}{2\sigma_0} \left\{ \int_{\rm (I)} (q+k)^2 d\sigma_g + \int_{\rm (II)} (\bar{q}+k)^2 d\sigma_g \right\} \; . \tag{C.11}$$

The virtual correction does not enter here, since the pure $q\bar{q}$ final state has jet masses equal to 0. The result of this simple computation leads to

$$\langle m^2_{\rm jet} \rangle = \frac{\alpha_s C_{\rm F}}{\pi} s \; . \tag{C.12}$$

Another interesting variable often used in experimental studies is the thrust T, defined by

$$T = \max_{\hat{T}} \sum_i |\vec{p}_i \cdot \hat{T}| \; / \; \sum_i |\vec{p}_i|$$

where $\hat{T}$ is the thrust axis, defined so as to maximize T. For three-body final states, $\hat{T}$ is the direction of the highest-energy parton, and T is proportional to twice its energy:

$$T = 2\frac{\bar{q}_0}{\sqrt{s}} \equiv 1 - \frac{(q+k)^2}{s} = 1 - \frac{m^2_{\rm jet}}{s} \; . \tag{C.13}$$

As a result,

$$\langle 1 - T \rangle = \frac{\alpha_s C_{\rm F}}{\pi} \; . \tag{C.14}$$

At LEP, $\langle 1-T \rangle \simeq \frac{0.120}{\pi} \times \frac{4}{3} \simeq 0.05$. The terms neglected in the soft-gluon approximation we used throughout can be calculated, and give some small correction to the above results. Corrections will likewise come from higher-order effects. State-of-the-art calculations exist which evaluate all these ‘shape variables’ (and more!) up to $\mathcal{O}(\alpha_s^2)$ accuracy, including a full next-to-leading-log accurate resummation of higher-order logarithms (such as the $\log 1/y$ terms we encountered in the discussion of jet rates, or terms of the form $\log^n(1-T)$ which appear at higher orders in the evaluation of the thrust distributions). These calculations allow a reliable estimate of several different observables directly proportional to α_s, and provide the theoretical input for the extraction of α_s from the LEP QCD data [8].

Note that non-perturbative corrections proportional to $\frac{\Lambda}{\sqrt{s}}$, with $\Lambda \sim 1$ GeV, can have a significant impact on the extraction of α_s. For example, a $\frac{\Lambda}{\sqrt{s}}$ correction to $\langle 1-T \rangle$ would be a 20% effect:

$$\frac{\Lambda}{\sqrt{s}} \sim 0.01 \; , \qquad \langle 1-T \rangle_{\rm PT} \simeq 0.05 \; .$$

Indeed one measures $\langle 1-T \rangle_{\rm LEP} = 0.068 \pm 0.003$, compared with the full perturbation theory QCD prediction of 0.055 (using $\alpha_s = 0.120$).

Acknowledgements

It is a pleasure to thank the organizers of this School for the successful efforts made to bring top-quality students together and to provide a great environment for physics discussions, and for a pleasant time as well.

Bibliography

Standard Textbooks

[1] R.P. Feynman, *Photon–Hadron Interactions* (W.A. Benjamin, NY, 1972).

[2] B.L. Ioffe, V.A. Khoze and L.N. Lipatov, *Hard Processes* (North Holland, 1984).

[3] T. Muta, *Foundations of QCD* (World Scientific, 1987).

[4] V. Barger and R.J.N. Phillips, *Collider Physics* (Addison Wesley, 1987).

[5] R. Field, *Applications of Perturbative QCD* (Addison Wesley, 1989).

[6] Yu.L. Dokshitzer, V.A. Khoze, A.H. Mueller and S.I. Troyan, *Basics of Perturbative QCD* (Editions Frontières, 1991).

[7] M.E. Peskin and D.V. Schroeder, *An Introduction to Quantum Field Theory* (Addison-Wesley, 1995).

[8] R.K. Ellis, W.J. Stirling and B.R. Webber, *QCD and Collider Physics* (Cambridge University Press, 1996).

Pedagogical Reviews

[9] P. Nason, lectures delivered at the 1997 European School of High-Energy Physics, Menstrup, Denmark, CERN 98-03 .

[10] Yu.L. Dokshitzer, lectures delivered at the 1995 European School of High-Energy Physics, Dubna, Russia, CERN 96-04.

[11] M. Neubert, lectures delivered at the 1995 European School of High-Energy Physics, Dubna, Russia, CERN 96-04.

Review Articles

[12] G. Altarelli, *Phys. Rep.* **81** (1982) 1.

[13] A.H. Mueller, *Phys. Rev.* **73** (1981) 237.

[14] Yu.L. Dokshitzer, D.I. Dyakonov and S.I. Trojan, *Phys. Rep.* **58** (1980) 270.

[15] A. Bassetto, M. Ciafaloni and G. Marchesini, *Phys. Rep.* **100** (1983) 201.

[16] M.L. Mangano and S.J. Parke, *Phys. Rep.* **200** (1991) 301.

Historical Reviews

[17] D.J. Gross, `hep-th/9809060`.

[18] G. 't Hooft, `hep-th/9808154`.

CP violation in meson decays

Y. Nir
Weizmann Institute of Science, Rehovot, Israel

Abstract

This contribution is aimed at graduate students in the field of (theoretical and experimental) high-energy physics. The main topics covered are (i) the flavour sector of the Standard Model and the Kobayashi–Maskawa mechanism of CP violation; (ii) formalism and theoretical interpretation of CP violation in meson decays; (iii) K decays; (iv) D decays; (v) B decays ($b \to c\bar{c}s$, $b \to s\bar{s}s$, $b \to u\bar{u}d$ and $b \to c\bar{u}s, u\bar{c}s$); and (vi) CP violation as a probe of new physics and, in particular, of supersymmetry.

1 Introduction

The Standard Model (SM) predicts that the only way that CP is violated is through the Kobayashi–Maskawa mechanism [1]. Specifically, the source of CP violation is a *single* phase in the mixing matrix that describes the charged-current weak interactions of quarks. In this introduction, we briefly review the present evidence that supports the Kobayashi–Maskawa (KM) picture of CP violation, as well as the various arguments against this picture.

1.1 Why believe the Kobayashi–Maskawa mechanism?

Experiments have measured to date nine independent CP-violating observables[1]:

1. Indirect CP violation in $K \to \pi\pi$ decays [2] and in $K \to \pi\ell\nu$ decays is given by
$$\varepsilon_K = (2.28 \pm 0.02) \times 10^{-3}\, e^{i\pi/4}\ . \tag{1}$$
2. Direct CP violation in $K \to \pi\pi$ decays [3–5] is given by
$$\varepsilon'/\varepsilon = (1.72 \pm 0.18) \times 10^{-3}\ . \tag{2}$$
3. CP violation in the interference of mixing and decay in the $B \to \psi K_S$ and other, related modes is given by [6, 7]:
$$S_{\psi K_S} = +0.69 \pm 0.03\ . \tag{3}$$
4. CP violation in the interference of mixing and decay in the $B \to K^+K^-K_S$ mode is given by [8,9]
$$S_{K^+K^-K_S} = -0.45 \pm 0.13\ . \tag{4}$$
5. CP violation in the interference of mixing and decay in the $B \to D^{*+}D^{*-}$ mode is given by [10, 11]
$$S_{D^{*+}D^{*-}} = -0.75 \pm 0.23\ . \tag{5}$$
6. CP violation in the interference of mixing and decay in the $B \to \eta' K^0$ modes is given by [12–14]
$$S_{\eta' K_S} = +0.50 \pm 0.09(0.13)\ . \tag{6}$$

[1] The list of measured observables in B decays is somewhat conservative. I include only observables where the combined significance of Babar and Belle measurements (taking an inflated error in case of inconsistencies) is above 3σ.

7. CP violation in the interference of mixing and decay in the $B \to f_0 K_S$ mode is given by [13,15]

$$S_{f_0 K_S} = -0.75 \pm 0.24 \; . \tag{7}$$

8. Direct CP violation in the $\overline{B}^0 \to K^- \pi^+$ mode is given by [16,17]

$$\mathcal{A}_{K^\mp \pi^\pm} = -0.115 \pm 0.018 \; . \tag{8}$$

9. Direct CP violation in the $B \to \rho\pi$ mode is given by [18,19]

$$\mathcal{A}_{\rho\pi}^{-+} = -0.48 \pm 0.14 \; . \tag{9}$$

All nine measurements — as well as many other, where CP violation is not (yet) observed at a level higher than 3σ — are consistent with the KM picture of CP violation. In particular, the measurement of the phase β from the CP asymmetry $B \to \psi K$ and the measurement of the phase α from CP asymmetries and decay rates in the $B \to \pi\pi, \rho\pi$ and $\rho\rho$ modes have provided the first two precision tests of CP violation in the SM. Since the model has passed these tests successfully, we are able, for the first time, to make the following statement: *The Kobayashi–Maskawa phase is, very likely, the dominant source of CP violation in low-energy flavour-changing processes.*

In contrast, various alternative scenarios of CP violation that have been phenomenologically viable for many years are now unambiguously excluded. Two important examples are the following.

– The superweak framework [20], that is, the idea that CP violation is purely indirect, is excluded by the evidence that $\varepsilon'/\varepsilon \neq 0$.
– Approximate CP, that is, the idea that all CP-violating phases are small (see, for example, [21]), is excluded by the evidence that $S_{\psi K_S} = \mathcal{O}(1)$.

Indeed, I am not aware of any viable, reasonably motivated, scenario which provides a complete alternative to the KM mechanism, that is, of a framework where the KM phase plays no significant role in the observed CP violation.

The experimental results from the B-factories, such as those in Eqs. (3)–(9), and their implications for theory signify a new era in the study of CP violation. In this contribution we explain these recent developments and their significance.

1.2 Why doubt the Kobayashi–Maskawa mechanism?

1.2.1 The baryon asymmetry of the universe

Baryogenesis is a consequence of CP-violating processes [22]. Therefore the present baryon number, which is accurately deduced from nucleosynthesis and Cosmic Microwave Background Radiation (CMBR) constraints,

$$Y_B \equiv \frac{n_B - n_{\overline{B}}}{s} \simeq 9 \times 10^{-11} \; , \tag{10}$$

is essentially a CP-violating observable! It can be added to the list of known CP-violating observables, Eqs. (1)–(9). Within a given model of CP violation, one can check for consistency between the data from cosmology, Eq. (10), and those from laboratory experiments.

The surprising point is that the Kobayashi–Maskawa mechanism for CP violation fails to account for Eq. (10). It predicts present baryon number density that is many orders of magnitude below the observed value [23–25]. This failure is independent of other aspects of the SM: The suppression of Y_B from CP violation is much too strong, even if the departure from thermal equilibrium is induced by mechanisms beyond the SM. This situation allows us to make the following statement: *There must exist sources of CP violation beyond the Kobayashi–Maskawa phase.*

Two important examples of viable models of baryogenesis are the following.

1. Leptogenesis [26]: a lepton asymmetry is induced by CP-violating decays of heavy fermions that are singlets of the SM gauge group (sterile neutrinos). Departure from thermal equilibrium is provided if the lifetime of the heavy neutrino is long enough that it decays when the temperature is below its mass. Processes that violate $B + L$ are fast before the electroweak phase transition and partially convert the lepton asymmetry into a baryon asymmetry. The CP-violating parameters may be related to CP violation in the mixing matrix for the light neutrinos (but this is a model-dependent issue [27]).
2. Electroweak baryogenesis (for a review see Ref. [28]): the source of the baryon asymmetry is the interactions of top (anti)quarks with the Higgs field during the electroweak phase transition. CP violation is induced, for example, by supersymmetric interactions. Sphaleron configurations provide baryon-number-violating interactions. Departure from thermal equilibrium is provided by the wall between the false vacuum ($\langle\phi\rangle = 0$) and the expanding bubble with the true vacuum, where electroweak symmetry is broken.

1.2.2 The strong CP problem

Nonperturbative QCD effects induce an additional term in the SM Lagrangian,

$$\mathcal{L}_\theta = \frac{\theta_{\rm QCD}}{32\pi^2}\epsilon_{\mu\nu\rho\sigma}F^{\mu\nu a}F^{\rho\sigma a} \ . \tag{11}$$

This term violates CP. In particular, it induces an electric dipole moment (EDM) to the neutron. The leading contribution in the chiral limit is given by [29]

$$d_N = \frac{g_{\pi NN}\bar{g}_{\pi NN}}{4\pi^2 M_N}\ln\frac{M_N}{m_\pi} \approx 5\times 10^{-16}\,\theta_{\rm QCD}\; e\ {\rm cm}\ , \tag{12}$$

where M_N is the nucleon mass, and $g_{\pi NN}$ ($\bar{g}_{\pi NN}$) is the pseudoscalar coupling (CP-violating scalar coupling) of the pion to the nucleon. (The leading contribution in the large N_c limit was calculated in the Skyrme model [30] and leads to a similar estimate.) The experimental bound on d_N is given by [31]

$$d_N \leq 6.3\times 10^{-26}\; e\ {\rm cm}\ . \tag{13}$$

It leads to the following bound on $\theta_{\rm QCD}$:

$$\theta_{\rm QCD} \lesssim 10^{-10}\ . \tag{14}$$

Since $\theta_{\rm QCD}$ arises from nonperturbative QCD effects, it is impossible to calculate it. Yet, there are good reasons to expect that these effects should yield $\theta_{\rm QCD} = \mathcal{O}(1)$ (for a review, see Ref. [32]). Within the SM, a value as small as in Eq. (14) is unnatural, since setting $\theta_{\rm QCD}$ to zero does not add symmetry to the model. [In particular, as we will see below, CP is violated by $\delta_{\rm KM} = \mathcal{O}(1)$.] Understanding why CP is so small in the strong interactions is the strong CP problem.

It seems then that the strong CP problem is a clue to new physics. Among the solutions that have been proposed are a massless u-quark (for a review, see Ref. [33]), the Peccei–Quinn mechanism [34,35] and spontaneous CP violation.

1.2.3 New physics

Almost any extension of the SM provides new sources of CP violation. For example, in the supersymmetric extension (with R-parity), there are 44 independent phases, most of them in flavour-changing couplings. If there is new physics at or below the TeV scale, it is quite likely that the KM phase is not the only source of CP violation that is playing a role in meson decays.

1.3 Will new CP violation be observed in experiments?

The SM picture of CP violation is testable because the Kobayashi–Maskawa mechanism is unique and predictive. These features are mainly related to the fact that there is a single phase that is responsible for all CP violation. As a consequence of this situation, one finds two classes of tests:

1. Correlations. Many independent CP-violating observables are correlated within the SM. For example, the SM predicts that the CP asymmetries in $B \to \psi K_S$ and in $B \to \phi K_S$, which proceed through different quark transitions, are equal to each other (to a few per cent accuracy) [36, 37]. Another important example is the strong SM correlation between CP violation in $B \to \psi K_S$ and in $K \to \pi\nu\bar{\nu}$ [38–40]. It is a significant fact, in this context, that several CP-violating observables can be calculated with very small hadronic uncertainties. To search for violations of the correlations, precise measurements are important.
2. Zeros. Since the KM phase appears in flavour-changing, weak-interaction couplings of quarks, and only if all three generations are involved, many CP-violating observables are predicted to be negligibly small. For example, the transverse lepton polarization in semileptonic meson decays, CP violation in $t\bar{t}$ production, tree-level D decays, and (assuming $\theta_{\rm QCD} = 0$) the electric dipole moment of the neutron are all predicted to be orders of magnitude below the (present and near future) experimental sensitivity. To search for lifted zeros, measurements of CP violation in many different systems should be performed.

The strongest argument that new sources of CP violation must exist in Nature comes from baryogenesis. Whether the CP violation that is responsible for baryogenesis would be manifest in measurements of CP asymmetries in B decays depends on two issues.

1. The scale of the new CP violation. If the relevant scale is very high, such as in leptogenesis, the effects cannot be signalled in these measurements. To estimate the limit on the scale, the following three facts are relevant: First, the SM contributions to CP asymmetries in B decays are $\mathcal{O}(1)$. Second, the expected experimental accuracy would reach in some cases the few per cent level. Third, the contributions from new physics are expected to be suppressed by $(\Lambda_{\rm EW}/\Lambda_{\rm NP})^2$. The conclusion is that, if the new source of CP violation is related to physics at $\Lambda_{\rm NP} \gg 1$ TeV, it cannot be signalled in B decays. Only if the true mechanism is electroweak baryogenesis, can it potentially affect B decays.
2. The flavour dependence of the new CP violation. If it is flavour diagonal, its effects on B decays would be highly suppressed. It can still manifest itself in other, flavour-diagonal CP-violating observables, such as electric dipole moments.

We conclude that new measurements of CP asymmetries in meson decays are particularly sensitive to new sources of CP violation that come from physics at (or below) the few TeV scale and that are related to flavour-changing couplings. This is, for example, the case, in certain supersymmetric models of baryogenesis [41,42]. The search for electric dipole moments can reveal the existence of new flavour-diagonal CP violation.

Of course, there could be new flavour physics at the TeV scale that is not related to the baryon asymmetry and may give signals in B decays. The best motivated extension of the SM where this situation is likely is that of supersymmetry.

Finally, we would like to mention that, in the past, flavour physics and the physics of CP violation did indeed lead to the discovery of new physics or to probing it before it was directly observed in experiments:

– the smallness of $\dfrac{\Gamma(K_L \to \mu^+\mu^-)}{\Gamma(K^+ \to \mu^+\nu)}$ led to predicting a fourth (the charm) quark;
– the size of Δm_K led to a successful prediction of the charm mass;

– the size of Δm_B led to a successful prediction of the top mass;
– the measurement of ε_K led to predicting the third generation.

2 The Kobayashi–Maskawa mechanism

2.1 Yukawa interactions are the source of CP violation

A model of elementary particles and their interactions is defined by three ingredients:

1. the symmetries of the Lagrangian;
2. the representations of fermions and scalars;
3. the pattern of spontaneous symmetry breaking.

The Standard Model (SM) is defined as follows:

1. The gauge symmetry is
$$G_{\rm SM} = SU(3)_{\rm C} \times SU(2)_{\rm L} \times U(1)_{\rm Y} \; . \tag{15}$$
2. There are three fermion generations, each consisting of five representations of $G_{\rm SM}$:
$$Q^I_{Li}(3,2)_{+1/6},\; U^I_{Ri}(3,1)_{+2/3},\; D^I_{Ri}(3,1)_{-1/3},\; L^I_{Li}(1,2)_{-1/2},\; E^I_{Ri}(1,1)_{-1} \; . \tag{16}$$
Our notations mean that, for example, left-handed quarks, Q^I_L, are triplets of $SU(3)_{\rm C}$, doublets of $SU(2)_{\rm L}$ and carry hypercharge $Y = +1/6$. The super-index I denotes interaction eigenstates. The sub-index $i = 1, 2, 3$ is the flavour (or generation) index. There is a single scalar representation,
$$\phi(1,2)_{+1/2} \; . \tag{17}$$
3. The scalar ϕ assumes a VEV,
$$\langle \phi \rangle = \begin{pmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{pmatrix} \; , \tag{18}$$
so that the gauge group is spontaneously broken,
$$G_{\rm SM} \to SU(3)_{\rm C} \times U(1)_{\rm EM} \; . \tag{19}$$

The SM Lagrangian, $\mathcal{L}_{\rm SM}$, is the most general renormalizable Lagrangian that is consistent with the gauge symmetry (15), the particle content (16), (17) and the pattern of spontaneous symmetry breaking (18). It can be divided in three parts:

$$\mathcal{L}_{\rm SM} = \mathcal{L}_{\rm kinetic} + \mathcal{L}_{\rm Higgs} + \mathcal{L}_{\rm Yukawa} \; . \tag{20}$$

Concerning the kinetic terms, to maintain gauge invariance one has to replace the derivative with a covariant derivative:

$$D^\mu = \partial^\mu + i g_s G^\mu_a L_a + i g W^\mu_b T_b + i g' B^\mu Y \; . \tag{21}$$

Here G^μ_a are the eight gluon fields, W^μ_b the three weak-interaction bosons and B^μ the single hypercharge boson. The L_a's are $SU(3)_{\rm C}$ generators (the 3×3 Gell-Mann matrices $\frac{1}{2}\lambda_a$ for triplets, 0 for singlets), the T_b's are $SU(2)_{\rm L}$ generators (the 2×2 Pauli matrices $\frac{1}{2}\tau_b$ for doublets, 0 for singlets), and the Y's are the $U(1)_{\rm Y}$ charges. For example, for the left-handed quarks Q^I_L, we have

$$\mathcal{L}_{\rm kinetic}(Q_L) = i\overline{Q^I_{Li}}\gamma_\mu \left(\partial^\mu + \frac{i}{2} g_s G^\mu_a \lambda_a + \frac{i}{2} g W^\mu_b \tau_b + \frac{i}{6} g' B^\mu \right) Q^I_{Li} \; , \tag{22}$$

while for the left-handed leptons L_L^I, we have

$$\mathcal{L}_{\rm kinetic}(L_L) = i\overline{L_{Li}^I}\gamma_\mu\left(\partial^\mu + \frac{i}{2}gW_b^\mu\tau_b - ig'B^\mu\right)L_{Li}^I\,. \tag{23}$$

These parts of the interaction Lagrangian are always CP conserving.

The Higgs potential, which describes the scalar self interactions, is given by

$$\mathcal{L}_{\rm Higgs} = \mu^2\phi^\dagger\phi - \lambda(\phi^\dagger\phi)^2\,. \tag{24}$$

For the SM scalar sector, where there is a single doublet, this part of the Lagrangian is also CP conserving. For extended scalar sectors, such as that of a two Higgs doublet model, $\mathcal{L}_{\rm Higgs}$ can be CP violating. Even in the case that it is CP symmetric, it may lead to spontaneous CP violation.

The quark Yukawa interactions are given by

$$-\mathcal{L}_{\rm Yukawa}^{\rm quarks} = Y_{ij}^d\overline{Q_{Li}^I}\phi D_{Rj}^I + Y_{ij}^u\overline{Q_{Li}^I}\tilde{\phi}U_{Rj}^I + {\rm h.c.}\,. \tag{25}$$

This part of the Lagrangian is, in general, CP violating. More precisely, CP is violated if and only if [43]

$$\mathcal{I}m(\det[Y^dY^{d\dagger}, Y^uY^{u\dagger}]) \neq 0\,. \tag{26}$$

An intuitive explanation of why CP violation is related to *complex* Yukawa couplings goes as follows. The hermiticity of the Lagrangian implies that $\mathcal{L}_{\rm Yukawa}$ has its terms in pairs of the form

$$Y_{ij}\overline{\psi_{Li}}\phi\psi_{Rj} + Y_{ij}^*\overline{\psi_{Rj}}\phi^\dagger\psi_{Li}\,. \tag{27}$$

A CP transformation exchanges the operators

$$\overline{\psi_{Li}}\phi\psi_{Rj} \leftrightarrow \overline{\psi_{Rj}}\phi^\dagger\psi_{Li}\,, \tag{28}$$

but leaves their coefficients, Y_{ij} and Y_{ij}^*, unchanged. This means that CP is a symmetry of $\mathcal{L}_{\rm Yukawa}$ if $Y_{ij} = Y_{ij}^*$.

The lepton Yukawa interactions are given by

$$-\mathcal{L}_{\rm Yukawa}^{\rm leptons} = Y_{ij}^e\overline{L_{Li}^I}\phi E_{Rj}^I + {\rm h.c.}\,. \tag{29}$$

It leads, as we will see in the next section, to charged lepton masses but predicts massless neutrinos. Recent measurements of the fluxes of atmospheric and solar neutrinos provide evidence for neutrino masses (for a review, see Ref. [44]). That means that $\mathcal{L}_{\rm SM}$ cannot be a complete description of Nature. The simplest way to allow for neutrino masses is to add dimension-five (and, therefore, non-renormalizable) terms, consistent with the SM symmetry and particle content:

$$-\mathcal{L}_{\rm Yukawa}^{\rm dim-5} = \frac{Y_{ij}^\nu}{M}L_iL_j\phi\phi + {\rm h.c.}\,. \tag{30}$$

The parameter M has dimension of mass. The dimensionless couplings Y_{ij}^ν are symmetric ($Y_{ij}^\nu = Y_{ji}^\nu$). We refer to the SM extended to include the terms $\mathcal{L}_{\rm Yukawa}^{\rm dim-5}$ of Eq. (30) as the ‘extended SM’ (ESM):

$$\mathcal{L}_{\rm ESM} = \mathcal{L}_{\rm kinetic} + \mathcal{L}_{\rm Higgs} + \mathcal{L}_{\rm Yukawa} + \mathcal{L}_{\rm Yukawa}^{\rm dim-5}\,. \tag{31}$$

The inclusion of non-renormalizable terms is equivalent to postulating that the SM is only a low-energy effective theory, and that new physics appears at the scale M.

How many independent CP-violating parameters are there in $\mathcal{L}_{\rm Yukawa}^{\rm quarks}$? Each of the two Yukawa matrices Y^q ($q = u, d$) is 3×3 and complex. Consequently, there are 18 real and 18 imaginary parameters in these matrices. Not all of them are, however, physical. One can think of the quark Yukawa couplings as spurions that break a global symmetry,

$$U(3)_Q \times U(3)_D \times U(3)_U \;\rightarrow\; U(1)_B \; . \tag{32}$$

This means that there is freedom to remove nine real and 17 imaginary parameters [the number of parameters in three 3×3 unitary matrices minus the phase related to $U(1)_B$]. We conclude that there are 10 quark flavour parameters: nine real ones and a single phase. This single phase is the source of CP violation in the quark sector.

How many independent CP-violating parameters are there in the lepton Yukawa interactions? The matrix Y^e is a general complex 3×3 matrix and depends, therefore, on nine real and nine imaginary parameters. The matrix Y^ν is symmetric and depends on six real and six imaginary parameters. Not all of these 15 real and 15 imaginary parameters are physical. One can think of the lepton Yukawa couplings as spurions that break (completely) a global symmetry,

$$U(3)_L \times U(3)_E \; . \tag{33}$$

This means that six real and 12 imaginary parameters are not physical. We conclude that there are 12 lepton flavour parameters: nine real ones and three phases. These three phases induce CP violation in the lepton sector.

2.2 CKM mixing is the (only!) source of CP violation in the quark sector

Upon the replacement $\mathcal{R}e(\phi^0) \rightarrow \dfrac{v + H^0}{\sqrt{2}}$ [see Eq. (18)], the Yukawa interactions (25) give rise to mass terms:

$$-\mathcal{L}_M^q = (M_d)_{ij}\overline{D_{Li}^I}D_{Rj}^I + (M_u)_{ij}\overline{U_{Li}^I}U_{Rj}^I + \text{h.c.} \; , \tag{34}$$

where

$$M_q = \frac{v}{\sqrt{2}} Y^q \; , \tag{35}$$

and we decomposed the $SU(2)_{\rm L}$ quark doublets into their components:

$$Q_{Li}^I = \begin{pmatrix} U_{Li}^I \\ D_{Li}^I \end{pmatrix} \; . \tag{36}$$

The mass basis corresponds, by definition, to diagonal mass matrices. We can always find unitary matrices V_{qL} and V_{qR} such that

$$V_{qL} M_q V_{qR}^\dagger = M_q^{\rm diag} \quad (q = u, d) \; , \tag{37}$$

with $M_q^{\rm diag}$ diagonal and real. The quark mass eigenstates are then identified as

$$q_{Li} = (V_{qL})_{ij} q_{Lj}^I \; , \qquad q_{Ri} = (V_{qR})_{ij} q_{Rj}^I \quad (q = u, d) \; . \tag{38}$$

The charged-current interactions for quarks (that is the interactions of the charged $SU(2)_{\rm L}$ gauge bosons $W_\mu^\pm = \frac{1}{\sqrt{2}}[W_\mu^1 \mp i W_\mu^2]$), which in the interaction basis are described by (22), have a complicated form in the mass basis:

$$-\mathcal{L}_{W^\pm}^q = \frac{g}{\sqrt{2}} \overline{u_{Li}} \gamma^\mu (V_{uL} V_{dL}^\dagger)_{ij} d_{Lj} W_\mu^+ + \text{h.c.} \; . \tag{39}$$

The unitary 3×3 matrix,

$$V = V_{uL} V_{dL}^\dagger \; , \quad (V V^\dagger = \mathbf{1}) \; , \tag{40}$$

is the Cabibbo–Kobayashi–Maskawa (CKM) *mixing matrix* for quarks [1, 45]. A unitary 3×3 matrix depends on nine parameters: three real angles and six phases.

The form of the matrix is not unique.

1. There is freedom in defining V in that we can permute between the various generations. This freedom is fixed by ordering the up quarks and the down quarks by their masses, i.e., $(u_1, u_2, u_3) \to (u, c, t)$ and $(d_1, d_2, d_3) \to (d, s, b)$. The elements of V are written as follows:
$$V = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} . \tag{41}$$

2. There is further freedom in the phase structure of V. Let us define P_q $(q = u, d)$ to be diagonal unitary (phase) matrices. Then, if instead of using V_{qL} and V_{qR} for the rotation (38) to the mass basis we use $\tilde{V}_{qL}$ and $\tilde{V}_{qR}$, which we define as $\tilde{V}_{qL} = P_q V_{qL}$ and $\tilde{V}_{qR} = P_q V_{qR}$, we still maintain a legitimate mass basis since $M_q^{\rm diag}$ remains unchanged by such transformations. However, V does change:
$$V \to P_u V P_d^* \,. \tag{42}$$
This freedom is fixed by demanding that V has the minimal number of phases. In the three-generation case V has a single phase. (There are five phase differences between the elements of P_u and P_d and, therefore, five of the six phases in the CKM matrix can be removed.) This is the Kobayashi–Maskawa phase $\delta_{\rm KM}$ which is the single source of CP violation in the quark sector of the SM [1].

As a result of the fact that V is not diagonal, the $W^\pm$ gauge bosons couple to quarks (mass eigenstates) of different generations. Within the SM, this is the only source of *flavour-changing* quark interactions.

2.3 The three phases in the lepton mixing matrix

The leptonic Yukawa interactions (29) and (30) give rise to mass terms:
$$-\mathcal{L}^\ell_M = (M_e)_{ij}\overline{e^I_{Li}}e^I_{Rj} + (M_\nu)_{ij}\nu^I_{Li}\nu^I_{Lj} + \text{h.c.}\,, \tag{43}$$
where
$$M_e = \frac{v}{\sqrt{2}}Y^e\,, \quad M_\nu = \frac{v^2}{2M}Y^\nu\,, \tag{44}$$
and we decomposed the $SU(2)_{\rm L}$ lepton doublets into their components:
$$L^I_{Li} = \begin{pmatrix} \nu^I_{Li} \\ e^I_{Li} \end{pmatrix} . \tag{45}$$

We can always find unitary matrices V_{eL} and V_ν such that
$$V_{eL}M_eM_e^\dagger V_{eL}^\dagger = {\rm diag}(m_e^2, m_\mu^2, m_\tau^2)\,, \quad V_\nu M_\nu^\dagger M_\nu V_\nu^\dagger = {\rm diag}(m_1^2, m_1^2, m_3^2)\,. \tag{46}$$
The charged-current interactions for leptons, which in the interaction basis are described by Eq. (23), have the following form in the mass basis:
$$-\mathcal{L}^\ell_{W^\pm} = \frac{g}{\sqrt{2}}\overline{e_{Li}}\gamma^\mu(V_{eL}V_\nu^\dagger)_{ij}\nu_{Lj}W^-_\mu + \text{h.c.}\,. \tag{47}$$
The unitary 3×3 matrix,
$$U = V_{eL}V_\nu^\dagger\,, \tag{48}$$
is the *lepton mixing matrix* [46]. As with the CKM matrix, the form of the lepton mixing matrix is not unique. But there are differences in choosing conventions.

1. We can permute between the various generations. This freedom is usually fixed in the following way. We order the charged leptons by their masses, i.e., $(e_1, e_2, e_3) \to (e, \mu, \tau)$. With regard to the neutrinos, one takes into account that the atmospheric and solar neutrino data imply that $\Delta m^2_{\rm atm} \gg \Delta m^2_{\rm sol}$. It follows that one of the neutrino mass eigenstates is separated in its mass from the other two, which have a smaller mass difference. The convention is to denote this separated state by ν_3. For the remaining two neutrinos, ν_1 and ν_2, the convention is to call the heavier state ν_2. In other words, the three mass eigenstates are defined by the following conventions:

$$|\Delta m^2_{3i}| \gg |\Delta m^2_{21}| \,, \quad \Delta m^2_{21} > 0 \,. \tag{49}$$

Note in particular that ν_3 can be either heavier ('normal hierarchy') or lighter ('inverted hierarchy') than $\nu_{1,2}$. The elements of U are written as follows:

$$U = \begin{pmatrix} U_{e1} & U_{e2} & U_{e3} \\ U_{\mu 1} & U_{\mu 2} & U_{\mu 3} \\ U_{\tau 1} & U_{\tau 2} & U_{\tau 3} \end{pmatrix} \,. \tag{50}$$

2. There is further freedom in the phase structure of U. One can change the charged lepton mass basis by the transformation $e_{(L,R)i} \to e'_{(L,R)i} = (P_e)_{ii} e_{(L,R)i}$, where P_e is a phase matrix. There is, however, no similar freedom to redefine the neutrino mass eigenstates: from Eq. (43) one learns that a transformation $\nu_L \to P_\nu \nu_L$ will introduce phases into the diagonal mass matrix. This is related to the Majorana nature of neutrino masses, assumed in Eq. (30). The allowed transformation modifies U:

$$U \to P_e U \,. \tag{51}$$

This freedom is fixed by demanding that U have the minimal number of phases. Out of six phases of a generic unitary 3×3 matrix, the multiplication by P_e can be used to remove three. We conclude that the three-generation U matrix has three phases. One of these is the analogue of the Kobayashi–Maskawa phase. It is the only source of CP violation in processes that conserve lepton number, such as neutrino flavour oscillations. The other two phases can affect lepton number changing processes.

With $U \neq \mathbf{1}$, the $W^\pm$ gauge bosons couple to leptons (mass eigenstates) of different generations. Within the ESM, this is the only source of *flavour-changing* lepton interactions.

2.4 The flavour parameters

Examining the quark mass basis, one can easily identify the flavour parameters. In the quark sector, we have six quark masses and four mixing parameters: three mixing angles and a single phase.

The fact that there are only three real and one imaginary physical parameters in V can be made manifest by choosing an explicit parametrization. For example, the standard parametrization [47], used by the particle data group, is given by

$$V = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta} & c_{23}c_{13} \end{pmatrix} , \tag{52}$$

where $c_{ij} \equiv \cos\theta_{ij}$ and $s_{ij} \equiv \sin\theta_{ij}$. The three $\sin\theta_{ij}$ are the three real mixing parameters while δ is the Kobayashi–Maskawa phase. Another, very useful, example is the Wolfenstein parametrization, where the four mixing parameters are (λ, A, ρ, η) with $\lambda = |V_{us}| = 0.22$ playing the role of an expansion parameter and η representing the CP-violating phase [48, 49]:

$$V = \begin{pmatrix} 1 - \frac{1}{2}\lambda^2 - \frac{1}{8}\lambda^4 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda + \frac{1}{2}A^2\lambda^5[1 - 2(\rho + i\eta)] & 1 - \frac{1}{2}\lambda^2 - \frac{1}{8}\lambda^4(1 + 4A^2) & A\lambda^2 \\ A\lambda^3[1 - (1 - \frac{1}{2}\lambda^2)(\rho + i\eta)] & -A\lambda^2 + \frac{1}{2}A\lambda^4[1 - 2(\rho + i\eta)] & 1 - \frac{1}{2}A^2\lambda^4 \end{pmatrix} . \tag{53}$$

Various parametrizations differ in the way that the freedom of phase rotation, Eq. (42), is used to leave a single phase in V. One can define, however, a CP-violating quantity in $V_{\rm CKM}$ that is independent of the parametrization [43]. This quantity, $J_{\rm CKM}$, is defined through

$$\mathcal{I}m(V_{ij}V_{kl}V_{il}^*V_{kj}^*) = J_{\rm CKM}\sum_{m,n=1}^{3}\epsilon_{ikm}\epsilon_{jln}\ , \quad (i,j,k,l=1,2,3)\ . \tag{54}$$

In terms of the explicit parametrizations given above, we have

$$J_{\rm CKM} = c_{12}c_{23}c_{13}^2 s_{12}s_{23}s_{13}\sin\delta \simeq \lambda^6 A^2\eta\ . \tag{55}$$

It is interesting to translate the condition (26) to the language of the flavour parameters in the mass basis. One finds that the following is a necessary and sufficient condition for CP violation in the quark sector of the SM (we define $\Delta m_{ij}^2 \equiv m_i^2 - m_j^2$):

$$\Delta m_{tc}^2\Delta m_{tu}^2\Delta m_{cu}^2\Delta m_{bs}^2\Delta m_{bd}^2\Delta m_{sd}^2 J_{\rm CKM} \neq 0\ . \tag{56}$$

Equation (56) puts the following requirements on the SM in order that it violates CP:

- within each quark sector, there should be no mass degeneracy;
- none of the three mixing angles should be zero or $\pi/2$;
- the phase should be neither 0 nor π.

With regard to the lepton sector of the ESM, the flavour parameters are the six lepton masses, and six mixing parameters: three mixing angles and three phases. One can parametrize U in a convenient way by factorizing it into $U = \hat{U}P$. Here P is a diagonal unitary matrix that depends on two phases, e.g., $P = {\rm diag}(e^{i\phi_1}, e^{i\phi_2}, 1)$, while $\hat{U}$ can be parametrized in the same way as in Eq. (52). The advantage of this parametrization is that for the purpose of analysing lepton number conserving processes and, in particular, neutrino flavour oscillations, the parameters of P are usually irrelevant and one can use the same Chau-Keung parametrization as is being used for V. (An alternative way to understand these statements is to use a single-phase mixing matrix and put the extra two phases in the neutrino mass matrix. Then it is obvious that the effects of these 'Majorana-phases' always appear in conjunction with a factor of the Majorana mass that is a lepton-number-violating parameter.) On the other hand, the Wolfenstein parametrization, Eq. (53), is inappropriate for the lepton sector: it assumes $|V_{23}| \ll |V_{12}| \ll 1$, which does not hold here.

In order that the CP-violating phase δ in $\hat{U}$ be physically meaningful, i.e., there be CP violation that is not related to lepton-number violation, a condition similar to Eq. (56) should hold:

$$\Delta m_{\tau\mu}^2\Delta m_{\tau e}^2\Delta m_{\mu e}^2\Delta m_{32}^2\Delta m_{31}^2\Delta m_{21}^2 J_\ell \neq 0\ . \tag{57}$$

2.5 The unitarity triangles

A very useful concept is that of the *unitarity triangles*. We focus on the quark sector, but analogous triangles can be defined in the lepton sector. The unitarity of the CKM matrix leads to various relations among the matrix elements, e.g.,

$$V_{ud}V_{us}^* + V_{cd}V_{cs}^* + V_{td}V_{ts}^* = 0\ , \tag{58}$$
$$V_{us}V_{ub}^* + V_{cs}V_{cb}^* + V_{ts}V_{tb}^* = 0\ , \tag{59}$$
$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0\ . \tag{60}$$

Each of these three relations requires the sum of three complex quantities to vanish and so can be geometrically represented in the complex plane as a triangle. These are the unitarity triangles, though the

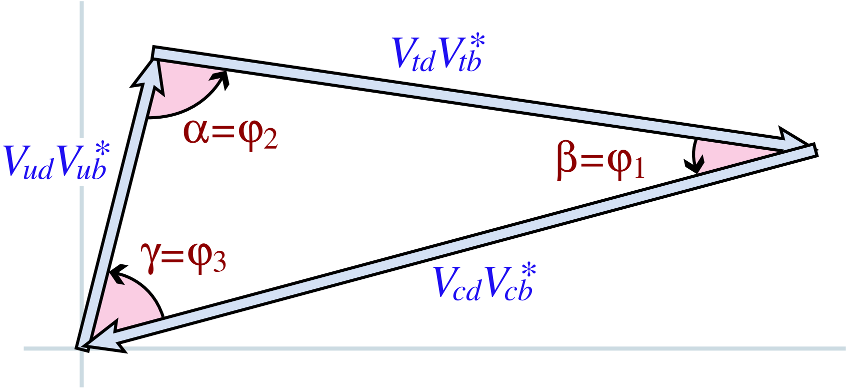


Fig. 1: Graphical representation of the unitarity constraint $V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0$ as a triangle in the complex plane

term 'unitarity triangle' is usually reserved for the relation (60) only. The unitarity triangle related to Eq. (60) is depicted in Fig. 1.

It is a surprising feature of the CKM matrix that all unitarity triangles are equal in area: the area of each unitarity triangle equals $|J_{\rm CKM}|/2$ and the sign of $J_{\rm CKM}$ gives the direction of the complex vectors around the triangles.

The rescaled unitarity triangle is derived from relation (60) by (a) choosing a phase convention such that $(V_{cd}V_{cb}^*)$ is real, and (b) dividing the lengths of all sides by $|V_{cd}V_{cb}^*|$. Step (a) aligns one side of the triangle with the real axis, and step (b) makes the length of this side 1. The form of the triangle is unchanged. Two vertices of the rescaled unitarity triangle are thus fixed at $(0,0)$ and $(1,0)$. The coordinates of the remaining vertex correspond to theter Wolfenstein parameters (ρ,η). The area of the rescaled unitarity triangle is $|\eta|/2$.

Depicting the rescaled unitarity triangle in the (ρ,η) plane, the lengths of the two complex sides are

$$R_u \equiv \left|\frac{V_{ud}V_{ub}}{V_{cd}V_{cb}}\right| = \sqrt{\rho^2+\eta^2}\ , \quad R_t \equiv \left|\frac{V_{td}V_{tb}}{V_{cd}V_{cb}}\right| = \sqrt{(1-\rho)^2+\eta^2}\ . \tag{61}$$

The three angles of the unitarity triangle are defined as follows [50, 51]:

$$\alpha \equiv \arg\left[-\frac{V_{td}V_{tb}^*}{V_{ud}V_{ub}^*}\right]\ , \quad \beta \equiv \arg\left[-\frac{V_{cd}V_{cb}^*}{V_{td}V_{tb}^*}\right]\ , \quad \gamma \equiv \arg\left[-\frac{V_{ud}V_{ub}^*}{V_{cd}V_{cb}^*}\right]\ . \tag{62}$$

They are physical quantities and can be independently measured by CP asymmetries in B decays. It is also useful to define the two small angles of the unitarity triangles (59) and (58):

$$\beta_s \equiv \arg\left[-\frac{V_{ts}V_{tb}^*}{V_{cs}V_{cb}^*}\right], \quad \beta_K \equiv \arg\left[-\frac{V_{cs}V_{cd}^*}{V_{us}V_{ud}^*}\right]\ . \tag{63}$$

To make predictions for CP-violating observables, we need to find the allowed ranges for the CKM phases. There are three ways to determine the CKM parameters (see, for example, Ref. [52]):

1. *Direct measurements* are related to SM tree-level processes. At present, we have direct measurements of $|V_{ud}|$, $|V_{us}|$, $|V_{ub}|$, $|V_{cd}|$, $|V_{cs}|$, $|V_{cb}|$ and $|V_{tb}|$.
2. *CKM unitarity* ($V^\dagger V = \mathbf{1}$) relates the various matrix elements. At present, these relations are useful to constrain $|V_{td}|$, $|V_{ts}|$, $|V_{tb}|$ and $|V_{cs}|$.
3. *Indirect measurements* are related to SM loop processes. At present, we constrain in this way $|V_{tb}V_{td}|$ (from Δm_B and Δm_{B_s}) and the phase structure of the matrix (for example, from ε_K and $S_{\psi K_S}$).

Direct measurements are expected to hold almost model independently. Most extensions of the SM have a special flavour structure that suppresses flavour-changing couplings and, in addition, have a mass scale $\Lambda_{\rm NP}$ that is higher than the electroweak breaking scale. Consequently, new physics contributions to tree-level processes are suppressed, compared to the SM ones, by at least $\mathcal{O}(m_Z^2/\Lambda_{\rm NP}^2) \ll 1$.

Unitarity holds if the only quarks (that is fermions in colour triplets with electric charge $+2/3$ or $-1/3$) are those of the three generations of the SM. This is the situation in many extensions of the SM, including the supersymmetric SM (SSM).

Using tree-level constraints and unitarity, the 90% confidence limits on the magnitude of the elements are [53]

$$\begin{pmatrix} 0.9739-0.9751 & 0.221\ \ -0.227 & 0.0029-0.0045 \\ 0.221\ \ -0.227 & 0.9730-0.9744 & 0.039\ \ -0.044 \\ 0.0048-0.014 & 0.037\ \ -0.043 & 0.9990-0.9992 \end{pmatrix} . \tag{64}$$

Note that $|V_{ub}|$ and $|V_{td}|$ are the only elements with uncertainties of order one.

Indirect measurements are sensitive to new physics. Take, for example, the $B^0-\overline{B}^0$ mixing amplitude. Within the SM, the leading contribution comes from an electroweak box diagram and is therefore $\mathcal{O}(g^4)$ and depends on small mixing angles, $(V_{td}^*V_{tb})^2$. (It is this dependence on the CKM elements that makes the relevant indirect measurements, particularly Δm_B and $S_{\psi K_S}$, very significant in improving our knowledge of the CKM matrix.) These suppression factors do not necessarily persist in extensions of the SM. For example, in the SSM there are (gluino-mediated) contributions of $\mathcal{O}(g_s^4)$ and the mixing angles could be comparable to or even larger than the SM ones. The validity of indirect measurements is then model dependent. Conversely, inconsistencies among indirect measurements (or between indirect and direct measurements) can give evidence for new physics.

When all available data are taken into account, one finds [54]

$$\begin{aligned} \lambda &= 0.226 \pm 0.001\ , \quad A = 0.83 \pm 0.03\ , & (65)\\ \bar\rho &= 0.21 \pm 0.04\ , \quad \bar\eta = 0.33 \pm 0.02\ , & (66)\\ \sin 2\beta &= 0.720 \pm 0.025\ , \quad \alpha = (99 \pm 7)^\circ\ , \quad \gamma = (58 \pm 7)^\circ\ , \quad \beta_s = (1.03 \pm 0.08)^\circ\ , & (67)\\ R_u &= 0.40 \pm 0.02\ , \quad R_t = 0.86 \pm 0.04\ . & (68) \end{aligned}$$

Of course, there are correlations between the various parameters. The present constraints on the shape of the unitarity triangle or, equivalently, the allowed region in the ρ–η plane, are presented in Fig. 2.

2.6 The uniqueness of the Standard Model picture of CP violation

In the previous subsections, we have learnt several features of CP violation as explained by the SM. It is important to understand that various reasonable (and often well-motivated) extensions of the SM provide examples where some or all of these features do not hold. Furthermore, until a few years ago, none of the special features of the Kobayashi–Maskawa mechanism of CP violation had been experimentally tested. This situation has dramatically changed recently. Let us survey some of the SM features, how they can be modified with new physics, and whether experiment has shed light on these questions.

1. $\delta_{\rm KM}$ *is the only source of CP violation in meson decays.* This is arguably the most distinguishing feature of the SM and gives the model a strong predictive power. It is violated in almost any low-energy extension. For example, in the supersymmetric extension of the SM there are 44 physical CP-violating phases, many of which affect meson decays. The measured value of $S_{\psi K_S}$ is consistent with the correlation between K and B decays that is predicted by the SM. The value of $S_{\phi K_S}$ is equal (within the present experimental accuracy) to that of $S_{\psi K_S}$, consistent with the

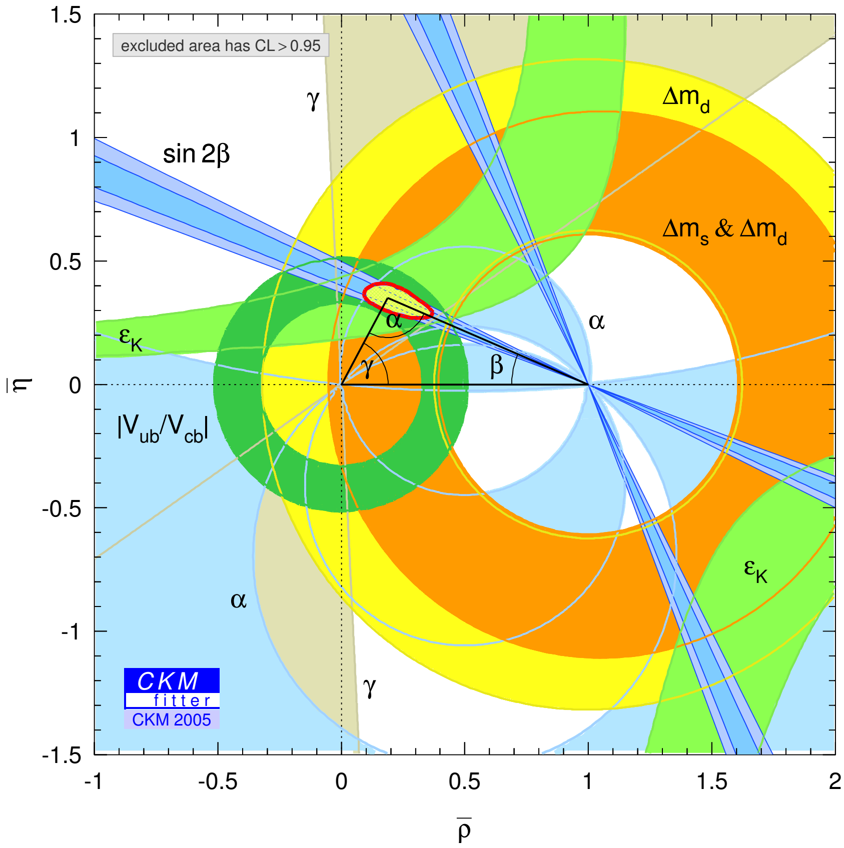


Fig. 2: Allowed region in the ρ–η plane. Superimposed are the individual constraints from charmless semileptonic B decays ($|V_{ub}/V_{cb}|$), mass differences in the B^0 (Δm_d) and B_s (Δm_s) neutral-meson systems, and CP violation in $K \to \pi\pi$ (ε_K), $B \to \psi K$ ($\sin 2\beta$), $B \to \pi\pi, \rho\pi, \rho\rho$ (α), and $B \to DK$ (γ). Taken from Ref. [54].

SM correlation between the asymmetries in $b \to s\bar{s}s$ and $b \to c\bar{c}s$ transitions. It is therefore very likely that $\delta_{\rm KM}$ is indeed the dominant source of CP violation in meson decays.

2. *CP violation is small in $K \to \pi\pi$ decays because of flavour suppression and not because CP is an approximate symmetry.* In many (though certainly not all) supersymmetric models, the flavour suppression is too mild, or entirely ineffective, requiring approximate CP to hold. The measurement of $S_{\psi K_S} = \mathcal{O}(1)$ confirms that not all CP-violating phases are small.
3. *CP violation appears in both $\Delta F = 1$ (decay) and $\Delta F = 2$ (mixing) amplitudes.* Superweak models suggest that CP is violated only in mixing amplitudes. The measurements of non-vanishing ε'/ε, $\mathcal{A}_{K^\mp\pi^\pm}$ and $\mathcal{A}_{\rho\pi}^{-+}$ confirm that there is CP violation in $\Delta S = 1$ and $\Delta B = 1$ processes.
4. *CP is not violated in the lepton sector.* Models that allow for neutrino masses, such as the ESM framework presented above, predict CP violation in leptonic charged-current interactions. Thus, while there is no measurement of leptonic CP violation, the data from neutrino oscillation experiments, which give evidence that neutrinos are massive and mix, make it very likely that charged-current weak interactions violate CP also in the lepton sector.
5. *CP violation appears only in the charged-current weak interactions and in conjunction with flavour-changing processes.* Here both various extensions of the SM (such as supersymmetry) and non-perturbative effects within the SM ($\theta_{\rm QCD}$) allow for CP violation in other types of interactions and in flavour-diagonal processes. In particular, it is difficult to avoid flavour-diagonal phases in the supersymmetric framework. The fact that no electric dipole moment has yet been measured poses difficulties for many models with diagonal CP violation (and, of course, is responsible for

the strong CP problem within the SM).

6. *CP is explicitly broken.* In various extensions of the scalar sector, it is possible to achieve spontaneous CP violation. It is very difficult to test this question experimentally.

This situation, where the SM has a unique and predictive description of CP violation, is the basis for the strong interest, experimental and theoretical, in CP violation.

3 Meson decays

The phenomenology of CP violation is superficially different in K, D, B, and B_s decays. This is primarily because each of these systems is governed by a different balance between decay rates, oscillations, and lifetime splitting. However, the underlying mechanisms of CP violation are identical for all pseudoscalar mesons.

In this section we present a general formalism for, and classification of, CP violation in the decay of a pseudoscalar meson P that might be a charged or neutral K, D, B, or B_s meson. Subsequent sections describe the CP-violating phenomenology, approximations, and alternate formalisms that are specific to each system. We follow here closely the discussion in Ref. [55].

3.1 Charged- and neutral-meson decays

We define decay amplitudes of a pseudoscalar meson P (which could be charged or neutral) and its CP conjugate $\overline{P}$ to a multi-particle final state f and its CP conjugate $\overline{f}$ as

$$A_f = \langle f|\mathcal{H}|P\rangle \quad , \quad \overline{A}_f = \langle f|\mathcal{H}|\overline{P}\rangle \quad , \quad A_{\overline{f}} = \langle \overline{f}|\mathcal{H}|P\rangle \quad , \quad \overline{A}_{\overline{f}} = \langle \overline{f}|\mathcal{H}|\overline{P}\rangle \; , \tag{69}$$

where $\mathcal{H}$ is the Hamiltonian governing weak interactions. The action of CP on these states introduces phases ξ_P and ξ_f that depend on their flavour content, according to

$$\begin{aligned} \text{CP}\,|P\rangle &= e^{+i\xi_P}\,|\overline{P}\rangle \quad , \quad \text{CP}\,|f\rangle = e^{+i\xi_f}\,|\overline{f}\rangle \; , \\ \text{CP}\,|\overline{P}\rangle &= e^{-i\xi_P}\,|P\rangle \quad , \quad \text{CP}\,|\overline{f}\rangle = e^{-i\xi_f}\,|f\rangle \; , \end{aligned} \tag{70}$$

so that (CP $)^2 = 1$. The phases ξ_P and ξ_f are arbitrary and unphysical because of the flavour symmetry of the strong interaction. If CP is conserved by the dynamics, [CP $, \mathcal{H}] = 0$, then A_f and $\overline{A}_{\overline{f}}$ have the same magnitude and an arbitrary unphysical relative phase

$$\overline{A}_{\overline{f}} = e^{i(\xi_f - \xi_P)}\, A_f \; . \tag{71}$$

3.2 Neutral-meson mixing

A state that is initially a superposition of P^0 and $\overline{P}^0$, say

$$|\psi(0)\rangle = a(0)|P^0\rangle + b(0)|\overline{P}^0\rangle \; , \tag{72}$$

will evolve in time acquiring components that describe all possible decay final states $\{f_1, f_2, \ldots\}$, that is,

$$|\psi(t)\rangle = a(t)|P^0\rangle + b(t)|\overline{P}^0\rangle + c_1(t)|f_1\rangle + c_2(t)|f_2\rangle + \cdots \; . \tag{73}$$

If we are interested in computing only the values of $a(t)$ and $b(t)$ (and not the values of all $c_i(t)$), and if the times t in which we are interested are much larger than the typical strong interaction scale, then we can use a much simplified formalism [56]. The simplified time evolution is determined by a 2×2 effective Hamiltonian $\mathcal{H}$ that is not Hermitian, since otherwise the mesons would only oscillate and not decay. Any complex matrix, such as $\mathcal{H}$, can be written in terms of Hermitian matrices M and Γ as

$$\mathcal{H} = M - \frac{i}{2}\Gamma \; . \tag{74}$$

M and Γ are associated with $(P^0, \overline{P}^0) \leftrightarrow (P^0, \overline{P}^0)$ transitions via off-shell (dispersive) and on-shell (absorptive) intermediate states, respectively. Diagonal elements of M and Γ are associated with the flavour-conserving transitions $P^0 \to P^0$ and $\overline{P}^0 \to \overline{P}^0$, while off-diagonal elements are associated with flavour-changing transitions $P^0 \leftrightarrow \overline{P}^0$.

The eigenvectors of $\mathcal{H}$ have well defined masses and decay widths. We introduce complex parameters $p_{L,H}$ and $q_{L,H}$ to specify the components of the strong interaction eigenstates, P^0 and $\overline{P}^0$, in the light (P_L) and heavy (P_H) mass eigenstates:

$$|P_{L,H}\rangle = p_{L,H}|P^0\rangle \pm q_{L,H}|\overline{P}^0\rangle \tag{75}$$

with the normalization $|p_{L,H}|^2 + |q_{L,H}|^2 = 1$. (Another possible choice, which is in standard usage for K mesons, defines the mass eigenstates according to their lifetimes: K_S for the short-lived and K_L for the long-lived state. The K_L is the heavier state.) If either CP or CPT is a symmetry of $\mathcal{H}$ (independently of whether T is conserved or violated) then $M_{11} = M_{22}$ and $\Gamma_{11} = \Gamma_{22}$, and solving the eigenvalue problem for $\mathcal{H}$ yields $p_L = p_H \equiv p$ and $q_L = q_H \equiv q$ with

$$\left(\frac{q}{p}\right)^2 = \frac{M_{12}^* - (i/2)\Gamma_{12}^*}{M_{12} - (i/2)\Gamma_{12}} \, . \tag{76}$$

If either CP or T is a symmetry of $\mathcal{H}$ (independently of whether CPT is conserved or violated), then M_{12} and Γ_{12} are relatively real, leading to

$$\left(\frac{q}{p}\right)^2 = e^{2i\xi_P} \quad \Rightarrow \quad \left|\frac{q}{p}\right| = 1 \, , \tag{77}$$

where ξ_P is the arbitrary unphysical phase introduced in Eq. (70). If, and only if, CP is a symmetry of $\mathcal{H}$ (independently of CPT and T) then both of the above conditions hold, with the result that the mass eigenstates are orthogonal:

$$\langle P_H|P_L\rangle = |p|^2 - |q|^2 = 0 \, . \tag{78}$$

From now on we assume that CPT is conserved.

The real and imaginary parts of the eigenvalues of $\mathcal{H}$ corresponding to $|P_{L,H}\rangle$ represent their masses and decay widths, respectively. The mass difference Δm and the width difference $\Delta\Gamma$ are defined as follows:

$$\Delta m \equiv M_H - M_L \, , \quad \Delta\Gamma \equiv \Gamma_H - \Gamma_L \, . \tag{79}$$

Note that here Δm is positive by definition, while the sign of $\Delta\Gamma$ is to be experimentally determined. (Alternatively, one can use the states defined by their lifetimes to have $\Delta\Gamma \equiv \Gamma_S - \Gamma_L$ positive by definition.) The average mass and width are given by

$$m \equiv \frac{M_H + M_L}{2} \, , \quad \Gamma \equiv \frac{\Gamma_H + \Gamma_L}{2} \, . \tag{80}$$

It is useful to define dimensionless ratios x and y:

$$x \equiv \frac{\Delta m}{\Gamma} \, , \quad y \equiv \frac{\Delta\Gamma}{2\Gamma} \, . \tag{81}$$

Solving the eigenvalue equation gives

$$(\Delta m)^2 - \frac{1}{4}(\Delta\Gamma)^2 = (4|M_{12}|^2 - |\Gamma_{12}|^2) \, , \quad \Delta m \Delta\Gamma = 4\mathcal{R}e(M_{12}\Gamma_{12}^*) \, . \tag{82}$$

3.3 CP-violating observables

All CP-violating observables in P and $\overline{P}$ decays to final states f and $\overline{f}$ can be expressed in terms of phase-convention-independent combinations of A_f, $\overline{A}_f$, $A_{\overline{f}}$ and $\overline{A}_{\overline{f}}$, together with, for neutral-meson decays only, q/p. CP violation in charged-meson decays depends only on the combination $|\overline{A}_{\overline{f}}/A_f|$, while CP violation in neutral-meson decays is complicated by $P^0 \leftrightarrow \overline{P}^0$ oscillations and depends, additionally, on $|q/p|$ and on $\lambda_f \equiv (q/p)(\overline{A}_f/A_f)$.

The decay rates of the two neutral K mass eigenstates, K_S and K_L, are different enough ($\Gamma_S/\Gamma_L \sim 500$) that one can, in most cases, actually study their decays independently. For neutral D, B, and B_s mesons, however, values of $\Delta\Gamma/\Gamma$ are relatively small and so both mass eigenstates must be considered in their evolution. We denote the state of an initially pure $|P^0\rangle$ or $|\overline{P}^0\rangle$ after an elapsed proper time t as $|P^0_{\rm phys}(t)\rangle$ or $|\overline{P}^0_{\rm phys}(t)\rangle$, respectively. Using the effective Hamiltonian approximation, we obtain

$$\begin{aligned} |P^0_{\rm phys}(t)\rangle &= g_+(t)\,|P^0\rangle - (q/p)\,g_-(t)|\overline{P}^0\rangle\,, \\ |\overline{P}^0_{\rm phys}(t)\rangle &= g_+(t)\,|\overline{P}^0\rangle - (p/q)\,g_-(t)|P^0\rangle\,, \end{aligned} \tag{83}$$

where

$$g_\pm(t) \equiv \frac{1}{2}\left(e^{-im_Ht-\frac{1}{2}\Gamma_Ht} \pm e^{-im_Lt-\frac{1}{2}\Gamma_Lt}\right)\,. \tag{84}$$

One obtains the following time-dependent decay rates:

$$\begin{aligned} \frac{d\Gamma[P^0_{\rm phys}(t)\to f]/dt}{e^{-\Gamma t}\mathcal{N}_f} &= \left(|A_f|^2+|(q/p)\overline{A}_f|^2\right)\cosh(y\Gamma t)+\left(|A_f|^2-|(q/p)\overline{A}_f|^2\right)\cos(x\Gamma t) \\ &+ 2\mathcal{R}e\big[(q/p)A_f^*\overline{A}_f\big]\sinh(y\Gamma t)-2\mathcal{I}m\big[(q/p)A_f^*\overline{A}_f\big]\sin(x\Gamma t)\,, \qquad (85) \\ \frac{d\Gamma[\overline{P}^0_{\rm phys}(t)\to f]/dt}{e^{-\Gamma t}\mathcal{N}_f} &= \left(|(p/q)A_f|^2+|\overline{A}_f|^2\right)\cosh(y\Gamma t)-\left(|(p/q)A_f|^2-|\overline{A}_f|^2\right)\cos(x\Gamma t) \\ &+ 2\mathcal{R}e\big[(p/q)A_f\overline{A}_f^*\big]\sinh(y\Gamma t)-2\mathcal{I}m\big[(p/q)A_f\overline{A}_f^*\big]\sin(x\Gamma t)\,, \qquad (86) \end{aligned}$$

where $\mathcal{N}_f$ is a common normalization factor. Decay rates to the CP-conjugate final state $\overline{f}$ are obtained analogously, with $\mathcal{N}_f = \mathcal{N}_{\overline{f}}$ and the substitutions $A_f \to A_{\overline{f}}$ and $\overline{A}_f \to \overline{A}_{\overline{f}}$ in Eqs. (85) and (86). Terms proportional to $|A_f|^2$ or $|\overline{A}_f|^2$ are associated with decays that occur without any net $P \leftrightarrow \overline{P}$ oscillation, while terms proportional to $|(q/p)\overline{A}_f|^2$ or $|(p/q)A_f|^2$ are associated with decays following a net oscillation. The $\sinh(y\Gamma t)$ and $\sin(x\Gamma t)$ terms of Eqs. (85) and (86) are associated with the interference between these two cases. Note that, in multi-body decays, amplitudes are functions of phase-space variables. Interference may be present in some regions but not others, and is strongly influenced by resonant substructure.

3.4 Classification of CP-violating effects

We distinguish three types of CP-violating effects in meson decays [57]:

[I] CP violation in decay is defined by

$$|\overline{A}_{\overline{f}}/A_f| \neq 1\,. \tag{87}$$

In charged meson decays, where mixing effects are absent, this is the only possible source of CP asymmetries:

$$\mathcal{A}_{f^\pm} \equiv \frac{\Gamma(P^-\to f^-)-\Gamma(P^+\to f^+)}{\Gamma(P^-\to f^-)+\Gamma(P^+\to f^+)} = \frac{|\overline{A}_{f^-}/A_{f^+}|^2-1}{|\overline{A}_{f^-}/A_{f^+}|^2+1}\,. \tag{88}$$

[II] CP violation in mixing is defined by

$$|q/p| \neq 1 \; . \tag{89}$$

In charged-current semileptonic neutral-meson decays $P, \overline{P} \to \ell^\pm X$ (taking $|A_{\ell^+X}| = |\overline{A}_{\ell^-X}|$ and $A_{\ell^-X} = \overline{A}_{\ell^+X} = 0$, as is the case in the SM, to lowest order, and in most of its reasonable extensions), this is the only source of CP violation, and can be measured via the asymmetry of 'wrong-sign' decays induced by oscillations:

$$\mathcal{A}_{\rm SL}(t) \equiv \frac{d\Gamma/dt[\overline{P}^0_{\rm phys}(t) \to \ell^+ X] - d\Gamma/dt[P^0_{\rm phys}(t) \to \ell^- X]}{d\Gamma/dt[\overline{P}^0_{\rm phys}(t) \to \ell^+ X] + d\Gamma/dt[P^0_{\rm phys}(t) \to \ell^- X]} = \frac{1-|q/p|^4}{1+|q/p|^4} \; . \tag{90}$$

Note that this asymmetry of time-dependent decay rates is actually time independent.

[III] CP violation in interference between a decay without mixing, $P^0 \to f$, and a decay with mixing, $P^0 \to \overline{P}^0 \to f$ (such an effect occurs only in decays to final states that are common to P^0 and $\overline{P}^0$, including all CP eigenstates), is defined by

$$\mathcal{I}m(\lambda_f) \neq 0 \; , \tag{91}$$

with

$$\lambda_f \equiv \frac{q}{p}\frac{\overline{A}_f}{A_f} \; . \tag{92}$$

This form of CP violation can be observed, for example, using the asymmetry of neutral-meson decays into final CP eigenstates $f_{\rm CP}$

$$\mathcal{A}_{f_{\rm CP}}(t) \equiv \frac{d\Gamma/dt[\overline{P}^0_{\rm phys}(t) \to f_{\rm CP}] - d\Gamma/dt[P^0_{\rm phys}(t) \to f_{\rm CP}]}{d\Gamma/dt[\overline{P}^0_{\rm phys}(t) \to f_{\rm CP}] + d\Gamma/dt[P^0_{\rm phys}(t) \to f_{\rm CP}]} \; . \tag{93}$$

If $\Delta\Gamma = 0$ and $|q/p| = 1$, as expected to a good approximation for B mesons but not for K mesons, then $\mathcal{A}_{f_{\rm CP}}$ has a particularly simple form [58–60]:

$$\begin{aligned} \mathcal{A}_f(t) &= S_f \sin(\Delta m t) - C_f \cos(\Delta m t) \; , \\ S_f &\equiv \frac{2\mathcal{I}m(\lambda_f)}{1+|\lambda_f|^2} \; , \quad C_f \equiv \frac{1-|\lambda_f|^2}{1+|\lambda_f|^2} \; . \end{aligned} \tag{94}$$

If, in addition, the decay amplitudes fulfil $|\overline{A}_{f_{\rm CP}}| = |A_{f_{\rm CP}}|$, the interference between decays with and without mixing is the only source of the asymmetry and

$$\mathcal{A}_{f_{\rm CP}}(t) = \mathcal{I}m(\lambda_{f_{\rm CP}}) \sin(x\Gamma t) \; . \tag{95}$$

4 Theoretical interpretation: general considerations

Consider the $P \to f$ decay amplitude A_f, and the CP-conjugate process, $\overline{P} \to \overline{f}$, with decay amplitude $\overline{A}_{\overline{f}}$. There are two types of phase that may appear in these decay amplitudes. Complex parameters in any Lagrangian term that contributes to the amplitude will appear in complex conjugate form in the CP-conjugate amplitude. Thus their phases appear in A_f and $\overline{A}_{\overline{f}}$ with opposite signs. In the SM, these phases occur only in the couplings of the $W^\pm$ bosons and hence are often called 'weak phases'. The weak phase of any single term is convention dependent. However, the difference between the weak phases in two different terms in A_f is convention independent. A second type of phase can appear in scattering or decay amplitudes even when the Lagrangian is real. Its origin is the possible contribution from intermediate on-shell states in the decay process. Since these phases are generated by CP-invariant interactions, they are the same in A_f and $\overline{A}_{\overline{f}}$. Usually the dominant rescattering is due to strong interactions, hence the designation ‘strong phases’ for the phase shifts so induced. Again, only the relative strong phases between different terms in the amplitude are physically meaningful.

The ‘weak’ and ‘strong’ phases discussed here appear in addition to the ‘spurious’ CP-transformation phases of Eq. (71). Those spurious phases are due to an arbitrary choice of phase convention, and do not originate from any dynamics or induce any CP violation. For simplicity, we set them to zero from here on.

It is useful to write each contribution a_i to A_f in three parts: its magnitude $|a_i|$, its weak phase ϕ_i, and its strong phase δ_i. If, for example, there are two such contributions, $A_f = a_1 + a_2$, we have

$$\begin{aligned} A_f &= |a_1|e^{i(\delta_1+\phi_1)} + |a_2|e^{i(\delta_2+\phi_2)} \ , \\ \overline{A}_{\overline{f}} &= |a_1|e^{i(\delta_1-\phi_1)} + |a_2|e^{i(\delta_2-\phi_2)} \ . \end{aligned} \tag{96}$$

Similarly, for neutral-meson decays, it is useful to write

$$M_{12} = |M_{12}|e^{i\phi_M} \quad , \quad \Gamma_{12} = |\Gamma_{12}|e^{i\phi_\Gamma} \ . \tag{97}$$

Each of the phases appearing in Eqs. (96) and (97) is convention dependent, but combinations such as $\delta_1 - \delta_2$, $\phi_1 - \phi_2$, $\phi_M - \phi_\Gamma$ and $\phi_M + \phi_1 - \overline{\phi}_1$ (where $\overline{\phi}_1$ is a weak phase contributing to $\overline{A}_f$) are physical.

It is now straightforward to evaluate the various asymmetries in terms of the theoretical parameters introduced here. We will do so with approximations that are often relevant to the most interesting measured asymmetries.

1. The CP asymmetry in charged-meson decays of Eq. (88) is given by

$$\mathcal{A}_{f^\pm} = -\frac{2|a_1 a_2|\sin(\delta_2-\delta_1)\sin(\phi_2-\phi_1)}{|a_1|^2+|a_2|^2+2|a_1 a_2|\cos(\delta_2-\delta_1)\cos(\phi_2-\phi_1)} \ . \tag{98}$$

   The quantity of most interest to theory is the weak phase difference $\phi_2 - \phi_1$. Its extraction from the asymmetry requires, however, that the amplitude ratio and the strong phase are known. Both quantities depend on non-perturbative hadronic parameters that are difficult to calculate.
2. In the approximation that $|\Gamma_{12}/M_{12}| \ll 1$ (valid for B and B_s mesons), the CP asymmetry in semileptonic neutral-meson decays, Eq. (90), is given by

$$\mathcal{A}_{\rm SL} = -\left|\frac{\Gamma_{12}}{M_{12}}\right| \sin(\phi_M - \phi_\Gamma). \tag{99}$$

   The quantity of most interest to theory is the weak phase $\phi_M - \phi_\Gamma$. Its extraction from the asymmetry requires, however, that $|\Gamma_{12}/M_{12}|$ is known. This quantity depends on long distance physics that is difficult to calculate.
3. In the approximations that only a single weak phase contributes to decay, $A_f = |a_f|e^{i(\delta_f+\phi_f)}$, and that $|\Gamma_{12}/M_{12}| = 0$, we obtain $|\lambda_f| = 1$ and the CP asymmetries in decays to a final CP eigenstate f, Eq. (93), with eigenvalue $\eta_f = \pm 1$ are given by

$$\mathcal{A}_{f_{\rm CP}}(t) = \mathcal{I}m(\lambda_f)\ \sin(\Delta m t) \text{ with } \mathcal{I}m(\lambda_f) = \eta_f \sin(\phi_M + 2\phi_f) \ . \tag{100}$$

   Note that the phase so measured is purely a weak phase, and no hadronic parameters are involved in the extraction of its value from $\mathcal{I}m(\lambda_f)$.

The discussion above allows us to introduce another classification.

1. **Direct CP violation** is one that cannot be accounted for by just $\phi_M \neq 0$. CP violation in decay (type I) belongs to this class.

2. **Indirect CP violation** is consistent with taking $\phi_M \neq 0$ and setting all other CP-violating phases to zero. CP violation in mixing (type II) belongs to this class.

With regard to type III CP violation, observing $\eta_{f_1}\mathcal{I}m(\lambda_{f_1}) \neq \eta_{f_2}\mathcal{I}m(\lambda_{f_2})$ (for the same decaying meson and two different final CP eigenstates f_1 and f_2) would establish direct CP violation. The significance of this classification is related to theory. In superweak models [20], CP violation appears only in diagrams that contribute to M_{12}, hence they predict that there is no direct CP violation. In most models and, in particular, in the SM, CP violation is both direct and indirect. The experimental observation of $\epsilon' \neq 0$ (see Section 5) excluded the superweak scenario.

5 K decays

CP violation was discovered in $K \to \pi\pi$ decays in 1964 [2]. The same mode provided the first evidence for direct CP violation [3–5].

The decay amplitudes actually measured in neutral K decays refer to the mass eigenstates K_L and K_S rather than to the K and $\overline{K}$ states referred to in Eq. (69). We define CP-violating amplitude ratios for two-pion final states,

$$\eta_{00} \equiv \frac{\langle\pi^0\pi^0|\mathcal{H}|K_L\rangle}{\langle\pi^0\pi^0|\mathcal{H}|K_S\rangle} \quad , \quad \eta_{+-} \equiv \frac{\langle\pi^+\pi^-|\mathcal{H}|K_L\rangle}{\langle\pi^+\pi^-|\mathcal{H}|K_S\rangle} \; . \tag{101}$$

Another important observable is the asymmetry of time-integrated semileptonic decay rates:

$$\delta_L \equiv \frac{\Gamma(K_L \to \ell^+\nu_\ell\pi^-) - \Gamma(K_L \to \ell^-\bar{\nu}_\ell\pi^+)}{\Gamma(K_L \to \ell^+\nu_\ell\pi^-) + \Gamma(K_L \to \ell^-\bar{\nu}_\ell\pi^+)} \; . \tag{102}$$

CP violation has been observed as an appearance of K_L decays to two-pion final states [53],

$$\begin{aligned} |\eta_{00}| &= (2.275 \pm 0.017) \times 10^{-3} & \phi_{00} &= 43.6^\circ \pm 0.8^\circ \\ |\eta_{+-}| &= (2.286 \pm 0.017) \times 10^{-3} & \phi_{+-} &= 43.2^\circ \pm 0.6^\circ \\ |\eta_{00}/\eta_{+-}| &= 0.9950 \pm 0.0008 & \phi_{00} - \phi_{+-} &= 0.4^\circ \pm 0.5^\circ \; , \end{aligned} \tag{103}$$

where ϕ_{ij} is the phase of the amplitude ratio η_{ij} determined without assuming CPT invariance. (A fit that assumes CPT gives [53] $\phi_{+-} = \phi_{00} = 43.49^\circ \pm 0.07^\circ$.) CP violation has also been observed in semileptonic K_L decays [53]:

$$\delta_L = (3.27 \pm 0.12) \times 10^{-3} \; , \tag{104}$$

where δ_L is a weighted average of muon and electron measurements, as well as in K_L decays to $\pi^+\pi^-\gamma$ and $\pi^+\pi^-e^+e^-$ [53].

Historically, CP violation in neutral K decays has been described in terms of parameters ϵ and ϵ'. The observables η_{00}, η_{+-}, and δ_L are related to these parameters, and to those of Section 3, by

$$\begin{aligned} \eta_{00} &= \tfrac{1-\lambda_{\pi^0\pi^0}}{1+\lambda_{\pi^0\pi^0}} &=& \quad \epsilon - 2\epsilon' \; , \\ \eta_{+-} &= \tfrac{1-\lambda_{\pi^+\pi^-}}{1+\lambda_{\pi^+\pi^-}} &=& \quad \epsilon + \epsilon' \; , \\ \delta_L &= \tfrac{1-|q/p|^2}{1+|q/p|^2} &=& \quad \frac{2\mathcal{R}e(\epsilon)}{1+|\epsilon|^2} \; , \end{aligned} \tag{105}$$

where, in the last line, we have assumed that $|A_{\ell^+\nu_\ell\pi^-}| = |\overline{A}_{\ell^-\bar{\nu}_\ell\pi^+}|$ and $|A_{\ell^-\bar{\nu}_\ell\pi^+}| = |\overline{A}_{\ell^+\nu_\ell\pi^-}| = 0$. A fit to the $K \to \pi\pi$ data yields [53]

$$|\epsilon| \;=\; (2.283 \pm 0.017) \times 10^{-3} \; ,$$

$$\mathcal{R}e(\epsilon'/\epsilon) \;=\; (1.67 \pm 0.26) \times 10^{-3}\,. \tag{106}$$

In discussing two-pion final states, it is useful to express the amplitudes $A_{\pi^0\pi^0}$ and $A_{\pi^+\pi^-}$ in terms of their isospin components via

$$\begin{aligned} A_{\pi^0\pi^0} &= \sqrt{\frac{1}{3}}|A_0|e^{i(\delta_0+\phi_0)} - \sqrt{\frac{2}{3}}|A_2|e^{i(\delta_2+\phi_2)}\,, \\ A_{\pi^+\pi^-} &= \sqrt{\frac{2}{3}}|A_0|e^{i(\delta_0+\phi_0)} + \sqrt{\frac{1}{3}}|A_2|e^{i(\delta_2+\phi_2)}\,, \end{aligned} \tag{107}$$

where we parametrize the amplitude $A_I(\overline{A}_I)$ for $K^0(\overline{K}^0)$ decay into two pions with total isospin $I = 0$ or 2 as

$$A_I \equiv \langle(\pi\pi)_I|\mathcal{H}|K^0\rangle = |A_I|e^{i(\delta_I+\phi_I)}\,, \quad \overline{A}_I \equiv \langle(\pi\pi)_I|\mathcal{H}|\overline{K}^0\rangle = |A_I|e^{i(\delta_I-\phi_I)}\,. \tag{108}$$

The smallness of $|\eta_{00}|$ and $|\eta_{+-}|$ allows us to approximate

$$\epsilon \simeq \frac{1}{2}(1-\lambda_{(\pi\pi)_{I=0}}), \qquad \epsilon' \simeq \frac{1}{6}\left(\lambda_{\pi^0\pi^0} - \lambda_{\pi^+\pi^-}\right)\,. \tag{109}$$

The parameter ϵ represents indirect CP violation, while ϵ' parametrizes direct CP violation: $\mathcal{R}e(\epsilon')$ measures CP violation in decay (type I), $\mathcal{R}e(\epsilon)$ measures CP violation in mixing (type II), and $\mathcal{I}m(\epsilon)$ and $\mathcal{I}m(\epsilon')$ measure the interference between decays with and without mixing (type III).

The following expressions for ϵ and ϵ' are useful for theoretical evaluations:

$$\epsilon \simeq \frac{e^{i\pi/4}}{\sqrt{2}}\frac{\mathcal{I}m(M_{12})}{\Delta m}\,, \qquad \epsilon' = \frac{i}{\sqrt{2}}\left|\frac{A_2}{A_0}\right| e^{i(\delta_2-\delta_0)}\sin(\phi_2-\phi_0)\,. \tag{110}$$

The expression for ϵ is only valid in a phase convention where $\phi_2 = 0$, corresponding to a real $V_{ud}V_{us}^*$, and in the approximation that also $\phi_0 = 0$. The phase of $\pi/4$ is approximate, and determined by hadronic parameters, $\arg\epsilon \approx \arctan(-2\Delta m/\Delta\Gamma)$, independently of the electroweak model. The calculation of ϵ benefits from the fact that $\mathcal{I}m(M_{12})$ is dominated by short distance physics. Consequently, the main source of uncertainty in theoretical interpretations of ϵ are the values of matrix elements such as $\langle K^0|(\bar{s}d)_{V-A}(\bar{s}d)_{V-A}|\overline{K}^0\rangle$. The expression for ϵ' is valid to first order in $|A_2/A_0| \sim 1/20$. The phase of ϵ' is experimentally determined, $\pi/2 + \delta_2 - \delta_0 \approx \pi/4$, and is independent of the electroweak model. Note that, accidentally, ϵ'/ϵ is real to a good approximation.

A future measurement of much interest is that of CP violation in the rare $K \to \pi\nu\bar{\nu}$ decays. The signal for CP violation is simply observing the $K_L \to \pi^0\nu\bar{\nu}$ decay. The effect here is that of interference between decays with and without mixing (type III) [61]:

$$\frac{\Gamma(K_L\to\pi^0\nu\bar{\nu})}{\Gamma(K^+\to\pi^+\nu\bar{\nu})} = \frac{1}{2}\left[1+|\lambda_{\pi\nu\bar{\nu}}|^2 - 2\,\mathcal{R}e(\lambda_{\pi\nu\bar{\nu}})\right] \simeq 1-\mathcal{R}e(\lambda_{\pi\nu\bar{\nu}})\,, \tag{111}$$

where in the last equation we neglect CP violation in decay and in mixing (expected, model independently, to be of order 10^{-5} and 10^{-3}, respectively). Such a measurement would be experimentally very challenging and theoretically very rewarding [62]. Similar to the CP asymmetry in $B \to J/\psi K_S$, the CP violation in $K \to \pi\nu\bar{\nu}$ decay is predicted to be large and can be very cleanly interpreted.

Within the SM, the $K_L \to \pi^0\nu\bar{\nu}$ decay is dominated by an intermediate top quark contribution and, consequently, can be cleanly interpreted in terms of CKM parameters [63]. (For the charged mode, $K^+ \to \pi^+\nu\bar{\nu}$, the contribution from an intermediate charm quark is not negligible and constitutes a source of hadronic uncertainty.) In particular, $\mathcal{B}(K_L \to \pi^0\nu\bar{\nu})$ provides a theoretically clean way to determine the Wolfenstein parameter η [64]:

$$\mathcal{B}(K_L\to\pi^0\nu\bar{\nu}) = \kappa_L X^2(m_t^2/m_W^2)A^4\eta^2\,, \tag{112}$$

where $\kappa_L = 1.80\times10^{-10}$ incorporates the value of the four-fermion matrix element which is deduced, using isospin relations, from $\mathcal{B}(K^+\to\pi^0e^+\nu)$, and $X(m_t^2/m_W^2)$ is a known function of the top mass.

5.1 Implications of ε_K

The measurement of ε_K has had (and still has) important implications. Two implications of historical importance are the following:

1. CP violation was discovered through the measurement of ε_K. Hence this measurement played a significant role in the history of particle physics.
2. The observation of $\varepsilon_K \neq 0$ led to the prediction that a third generation must exist, so that CP is violated in the SM. This provides an excellent example of how precision measurements at low energy can lead to the discovery of new physics (even if, at present, this new physics is old…)

Within the SM, $\mathcal{I}m(M_{12})$ is accounted for by box diagrams:

$$\varepsilon_K = e^{i\pi/4} C_\varepsilon B_K \mathcal{I}m(V_{ts}^* V_{td}) \left\{ \mathcal{R}e(V_{cs}^* V_{cd})[\eta_1 S_0(x_c) - \eta_3 S_0(x_c, x_t)] - \mathcal{R}e(V_{ts}^* V_{td}) \eta_2 S_0(x_t) \right\} , \tag{113}$$

where $C_\varepsilon \equiv \frac{G_F^2 f_K^2 m_K m_W^2}{6\sqrt{2}\pi^2 \Delta m_K}$ is a well known parameter, the η_i are QCD correction factors, S_0 is a kinematic factor, and B_K is the ratio between the matrix element of the four-quark operator and its value in the vacuum insertion approximation. The measurement of ε_K has the following implications within the SM.

- This measurement allowed one to set the value of δ_{KM}. Furthermore, by implying that $\delta_{\mathrm{KM}} = \mathcal{O}(1)$, it made the KM mechanism plausible. Having been the single measured CP-violating parameter it could not, however, serve as a test of the KM mechanism. More precisely, a value of $|\varepsilon_K| \gg 10^{-3}$ would have invalidated the KM mechanism, but any value $|\varepsilon_K| \lesssim 10^{-3}$ was acceptable. It is only the combination of the new measurements in B decays (particularly $S_{\psi K_S}$) with ε_K that provides the first precision test of the KM mechanism.
- Within the SM, the smallness of ε_K is not related to suppression of CP violation but rather to suppression of flavour violation. Specifically, it is the smallness of the ratio $|(V_{td}V_{ts})/(V_{ud}V_{us})| \sim \lambda^4$ that explains $|\varepsilon_K| \sim 10^{-3}$.
- Until recently, the measured value of ε_K provided a unique type of information on the KM phase. For example, the measurement of $\mathcal{R}e(\varepsilon_K) > 0$ tells us that $\eta > 0$ and excludes the lower half of the ρ–η plane. Such information cannot be obtained from any CP-conserving observable.
- The ε_K constraint in Eq. (113) gives hyperbole in the ρ–η plane. It is shown in Fig. 2. The measured value is consistent with all other CKM-related measurements and further narrows the allowed region.

Beyond the SM, ε_K is an extremely powerful probe of new physics. This aspect will be discussed later.

6 D decays

Unlike the case of neutral K, B, and B_s mixing, D^0–$\overline{D}^0$ mixing has not yet been observed. Long-distance contributions make it difficult to calculate the SM prediction for the D^0–$\overline{D}^0$ mixing parameters. Therefore, the goal of the search for D^0–$\overline{D}^0$ mixing is not to constrain the CKM parameters but rather to probe new physics. Here CP violation plays an important role [65]. Within the SM, the CP-violating effects are predicted to be negligibly small since the mixing and the relevant decays are described, to an excellent approximation, by physics of the first two generations. Observation of CP violation in D^0–$\overline{D}^0$ mixing (at a level much higher than $\mathcal{O}(10^{-3})$) will constitute an unambiguous signal of new physics.[2] At present, the most sensitive searches involve the $D \to K^+K^-$ and $D \to K^\pm\pi^\mp$ modes.

[2] In contrast, neither $y_D \sim 10^{-2}$ [66], nor $x_D \sim 10^{-2}$ [67] can be considered as evidence for new physics.

The neutral D mesons decay via a singly-Cabibbo-suppressed transition to the CP eigenstate K^+K^-. Since the decay proceeds via a SM tree diagram, it is very likely unaffected by new physics and, furthermore, dominated by a single weak phase. It is safe then to assume that direct CP violation plays no role here [68, 69]. In addition, given the experimental bounds [53], $x \equiv \Delta m/\Gamma \lesssim 0.03$ and $y \equiv \Delta\Gamma/(2\Gamma) = 0.008 \pm 0.005$, we can expand the decay rates to first order in these parameters. Using Eq. (85) with these assumptions and approximations yields, for $xt, yt \lesssim \Gamma^{-1}$,

$$\begin{aligned}\Gamma[D^0_{\rm phys}(t) \to K^+K^-] &= e^{-\Gamma t}|A_{KK}|^2[1 - |q/p|(y\cos\phi_D - x\sin\phi_D)\Gamma t]\ ,\\ \Gamma[\overline{D}{}^0_{\rm phys}(t) \to K^+K^-] &= e^{-\Gamma t}|A_{KK}|^2[1 - |p/q|(y\cos\phi_D + x\sin\phi_D)\Gamma t]\ ,\end{aligned} \tag{114}$$

where ϕ_D is defined via $\lambda_{K^+K^-} = -|q/p|e^{i\phi_D}$. [In the limit of CP conservation, choosing $\phi_D = 0$ is equivalent to defining the mass eigenstates by their CP eigenvalue: $|D_\mp\rangle = p|D^0\rangle \pm q|\overline{D}{}^0\rangle$, with $D_-(D_+)$ being the CP-odd (CP-even) state; that is, the state that does not (does) decay into K^+K^-.] Given the small values of x and y, the time dependences of the rates in Eq. (114) can be recast into purely exponential forms, but with modified decay-rate parameters [70]:

$$\begin{aligned}\Gamma_{D^0\to K^+K^-} &= \Gamma \times [1 + |q/p|(y\cos\phi_D - x\sin\phi_D)]\ ,\\ \Gamma_{\overline{D}{}^0\to K^+K^-} &= \Gamma \times [1 + |p/q|(y\cos\phi_D + x\sin\phi_D)]\ .\end{aligned} \tag{115}$$

One can define CP-conserving and CP-violating combinations of these two observables (normalized to the true width Γ):

$$\begin{aligned} Y &\equiv \frac{\Gamma_{\overline{D}{}^0\to K^+K^-} + \Gamma_{D^0\to K^+K^-}}{2\Gamma} - 1\\ &= \frac{|q/p| + |p/q|}{2}\, y\cos\phi_D - \frac{|q/p| - |p/q|}{2}\, x\sin\phi_D\ ,\\ \Delta Y &\equiv \frac{\Gamma_{\overline{D}{}^0\to K^+K^-} - \Gamma_{D^0\to K^+K^-}}{2\Gamma}\\ &= \frac{|q/p| + |p/q|}{2}\, x\sin\phi_D - \frac{|q/p| - |p/q|}{2}\, y\cos\phi_D\ .\end{aligned} \tag{116}$$

In the limit of CP conservation (and, in particular, within the SM), $Y = y$ and $\Delta Y = 0$.

The $K^\pm\pi^\mp$ states are not CP eigenstates but they are still common final states for D^0 and $\overline{D}{}^0$ decays. Since $D^0(\overline{D}{}^0) \to K^-\pi^+$ is a Cabibbo-favoured (doubly-Cabibbo-suppressed) process, these processes are particularly sensitive to x and/or $y = \mathcal{O}(\lambda^2)$. Taking into account that $|\lambda_{K^-\pi^+}|, |\lambda^{-1}_{K^+\pi^-}| \ll 1$ and $x, y \ll 1$, assuming that there is no direct CP violation (again, these are SM tree-level decays dominated by a single weak phase) and expanding the time-dependent rates for $xt, yt \lesssim \Gamma^{-1}$, one obtains

$$\begin{aligned}\frac{\Gamma[D^0_{\rm phys}(t) \to K^+\pi^-]}{\Gamma[\overline{D}{}^0_{\rm phys}(t) \to K^+\pi^-]} &= r_d^2 + r_d\left|\frac{q}{p}\right|(y'\cos\phi_D - x'\sin\phi_D)\Gamma t + \left|\frac{q}{p}\right|^2\frac{y^2+x^2}{4}(\Gamma t)^2\ ,\\ \frac{\Gamma[\overline{D}{}^0_{\rm phys}(t) \to K^-\pi^+]}{\Gamma[D^0_{\rm phys}(t) \to K^-\pi^+]} &= r_d^2 + r_d\left|\frac{p}{q}\right|(y'\cos\phi_D + x'\sin\phi_D)\Gamma t + \left|\frac{p}{q}\right|^2\frac{y^2+x^2}{4}(\Gamma t)^2\ ,\end{aligned} \tag{117}$$

where

$$\begin{aligned} y' &\equiv y\cos\delta - x\sin\delta\ ,\\ x' &\equiv x\cos\delta + y\sin\delta\ .\end{aligned} \tag{118}$$

The weak phase ϕ_D is the same as that of Eq. (114) (a consequence of the absence of direct CP violation), δ is a strong phase difference for these processes, and $r_d = \mathcal{O}(\tan^2\theta_c)$ is the amplitude ratio, $r_d = |\overline{A}_{K^-\pi^+}/A_{K^-\pi^+}| = |A_{K^+\pi^-}/\overline{A}_{K^+\pi^-}|$, that is, $\lambda_{K^-\pi^+} = r_d(q/p)e^{-i(\delta-\phi_D)}$ and $\lambda^{-1}_{K^+\pi^-} = r_d(p/q)e^{-i(\delta+\phi_D)}$. By fitting to the six coefficients of the various time dependences, one can extract r_d, $|q/p|$, (x^2+y^2), $y'\cos\phi_D$, and $x'\sin\phi_D$. In particular, finding CP violation, that is, $|q/p| \neq 1$ and/or $\sin\phi_D \neq 0$, would constitute evidence for new physics.

7 B decays

The upper bound on the CP asymmetry in semileptonic B decays [53] implies that CP violation in B^0–$\overline{B}^0$ mixing is a small effect, where we use $\mathcal{A}_{\rm SL}/2 \approx 1 - |q/p|$, see Eq. (90):

$$\mathcal{A}_{\rm SL} = (0.3 \pm 1.3) \times 10^{-2} \quad \Longrightarrow \quad |q/p| = 0.998 \pm 0.007\ . \tag{119}$$

The SM prediction is

$$\mathcal{A}_{\rm SL} = \mathcal{O}\left(\frac{m_c^2}{m_t^2}\sin\beta\right) \lesssim 0.001\ . \tag{120}$$

In models where Γ_{12}/M_{12} is approximately real, such as the SM, an upper bound on $\Delta\Gamma/\Delta m \approx \mathcal{R}e(\Gamma_{12}/M_{12})$ provides yet another upper bound on the deviation of $|q/p|$ from one. This constraint does not hold if Γ_{12}/M_{12} is approximately imaginary.

The small deviation (less than one per cent) of $|q/p|$ from one implies that, at the present level of experimental precision, CP violation in B mixing is a negligible effect. Thus, for the purpose of analysing CP asymmetries in hadronic B decays, we can use

$$\lambda_f = e^{-i\phi_B}(\overline{A}_f/A_f)\ , \tag{121}$$

where ϕ_B refers to the phase of M_{12}, see Eq. (97). Within the SM, the corresponding phase factor is given by

$$e^{-i\phi_B} = (V_{tb}^* V_{td})/(V_{tb}V_{td}^*)\ . \tag{122}$$

Some of the most interesting decays involve final states that are common to B^0 and $\overline{B}^0$ [71–73]. Here Eq. (94) applies [58–60]. The processes of interest proceed via quark transitions of the form $\bar{b} \to \bar{q}q\bar{q}'$ with $q' = s$ or d. For $q = c$ or u, there are contributions from both tree (t) and penguin (p^{q_u}, where $q_u = u, c, t$ is the quark in the loop) diagrams (see Fig. 3) which carry different weak phases:

$$A_f = \left(V_{qb}^* V_{qq'}\right) t_f + \sum_{q_u = u,c,t} \left(V_{q_u b}^* V_{q_u q'}\right) p_f^{q_u}\ . \tag{123}$$

(The distinction between tree and penguin contributions is a heuristic one, the separation by the operator that enters is more precise. For a detailed discussion of the more complete operator product approach, which also includes higher order QCD corrections, see, for example, Ref. [74].) Using CKM unitarity, these decay amplitudes can always be written in terms of just two CKM combinations. For example, for $f = \pi\pi$, which proceeds via $\bar{b} \to \bar{u}u\bar{d}$ transition, we can write

$$A_{\pi\pi} = (V_{ub}^* V_{ud})\, T_{\pi\pi} + (V_{tb}^* V_{td})\, P_{\pi\pi}^t\ , \tag{124}$$

where $T_{\pi\pi} = t_{\pi\pi} + p_{\pi\pi}^u - p_{\pi\pi}^c$ and $P_{\pi\pi}^t = p_{\pi\pi}^t - p_{\pi\pi}^c$. CP-violating phases in Eq. (124) appear only in the CKM elements, so that

$$\frac{\overline{A}_{\pi\pi}}{A_{\pi\pi}} = \frac{(V_{ub}V_{ud}^*)\, T_{\pi\pi} + (V_{tb}V_{td}^*)\, P_{\pi\pi}^t}{\left(V_{ub}^* V_{ud}\right) T_{\pi\pi} + \left(V_{tb}^* V_{td}\right) P_{\pi\pi}^t}\ . \tag{125}$$

For $f = J/\psi K$, which proceeds via $\bar{b} \to \bar{c}c\bar{s}$ transition, we can write

$$A_{\psi K} = (V_{cb}^* V_{cs})\, T_{\psi K} + (V_{ub}^* V_{us})\, P_{\psi K}^u\ , \tag{126}$$

where $T_{\psi K} = t_{\psi K} + p_{\psi K}^c - p_{\psi K}^t$ and $P_{\psi K}^u = p_{\psi K}^u - p_{\psi K}^t$. A subtlety arises in this decay that is related to the fact that $B^0 \to J/\psi K^0$ and $\overline{B}^0 \to J/\psi \overline{K}^0$. A common final state, e.g., $J/\psi K_S$, is reached only via K^0–$\overline{K}^0$ mixing. Consequently, the phase factor corresponding to neutral K mixing, $e^{-i\phi_K} = (V_{cd}^* V_{cs})/(V_{cd}V_{cs}^*)$, plays a role:

$$\frac{\overline{A}_{\psi K_S}}{A_{\psi K_S}} = -\frac{(V_{cb}V_{cs}^*)\, T_{\psi K} + (V_{ub}V_{us}^*)\, P_{\psi K}^u}{\left(V_{cb}^* V_{cs}\right) T_{\psi K} + \left(V_{ub}^* V_{us}\right) P_{\psi K}^u} \times \frac{V_{cd}^* V_{cs}}{V_{cd}V_{cs}^*}\ . \tag{127}$$

For $q = s$ or d, there are only penguin contributions to A_f, that is, $t_f = 0$ in Eq. (123). (The tree $\bar{b} \to \bar{u}u\bar{q}'$ transition followed by $\bar{u}u \to \bar{q}q$ rescattering is included below in the P^u terms.) Again, CKM unitarity allows us to write A_f in terms of two CKM combinations. For example, for $f = \phi K_S$, which proceeds via $\bar{b} \to \bar{s}s\bar{s}$ transition, we can write

$$\frac{\overline{A}_{\phi K_S}}{A_{\phi K_S}} = -\frac{(V_{cb}V_{cs}^*)\,P_{\phi K}^c + (V_{ub}V_{us}^*)\,P_{\phi K}^u}{(V_{cb}^*V_{cs})\,P_{\phi K}^c + (V_{ub}^*V_{us})\,P_{\phi K}^u} \times \frac{V_{cd}^*V_{cs}}{V_{cd}V_{cs}^*}\,, \tag{128}$$

where $P_{\phi K}^c = p_{\phi K}^c - p_{\phi K}^t$ and $P_{\phi K}^u = p_{\phi K}^u - p_{\phi K}^t$.

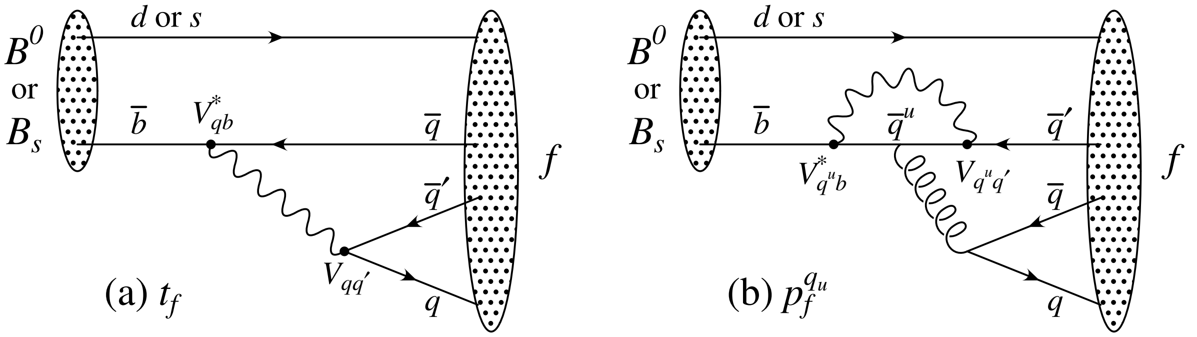


Fig. 3: Feynman diagrams for (a) tree and (b) penguin amplitudes contributing to $B^0 \to f$ or $B_s \to f$ via a $\bar{b} \to \bar{q}q\bar{q}'$ quark-level process

Since the amplitude A_f involves two different weak phases, the corresponding decays can exhibit both CP violation in the interference of decays with and without mixing, $S_f \neq 0$, and CP violation in decays, $C_f \neq 0$. [At the present level of experimental precision, the contribution to C_f from CP violation in mixing is negligible, see Eq. (119).] If the contribution from a second weak phase is suppressed, then the interpretation of S_f in terms of Lagrangian CP-violating parameters is clean, while C_f is small. If such a second contribution is not suppressed, S_f depends on hadronic parameters and, if the relevant strong phase is large, C_f is large.

A summary of $\bar{b} \to \bar{q}q\bar{q}'$ modes with $q' = s$ or d is given in Table 1. The $\bar{b} \to \bar{d}d\bar{q}$ transitions lead to final states that are similar to the $\bar{b} \to \bar{u}u\bar{q}$ transitions and have similar phase dependence. Final states that consist of two vector-mesons ($\psi\phi$ and $\phi\phi$) are not CP eigenstates, and angular analysis is needed to separate the CP-even from the CP-odd contributions.

The cleanliness of the theoretical interpretation of S_f can be assessed from the information in the last column of Table 1. In case of small uncertainties, the expression for S_f in terms of CKM phases can be deduced from the fourth column of Table 1 in combination with Eq. (122) [and, for $b \to q\bar{q}s$ decays, the example in Eq. (127)]. In the next three sections, we consider three interesting classes.

For B_s decays, one has to replace Eq. (122) with

$$e^{-i\phi_{B_s}} = (V_{tb}^*V_{ts})/(V_{tb}V_{ts}^*)\,. \tag{129}$$

Note that one expects $\Delta\Gamma_s/\Gamma_s = \mathcal{O}(0.1)$, and therefore y_{B_s} should not be put to zero in the expressions for the time-dependent decay rates, but $|q/p| = 1$ is expected to hold to an even better approximation than for B mesons. The CP asymmetry in $B_s \to D_s^+D_s^-$ (or in $B_s \to \psi\phi$ with angular analysis to disentangle the CP-even and CP-odd components of the final state) will determine $\sin 2\beta_s$, where β_s is defined in Eq. (63). Since the SM prediction is that this asymmetry is small, see Eq. (67), $\sin 2\beta_s \sim 0.036$, an observation of a $S_{B_s\to D_s^+D_s^-} \gg 0.04$ will provide evidence for new physics.

Table 1: Summary of $\bar{b} \to \bar{q}q\bar{q}'$ modes with $q' = s$ or d. The second and third columns give examples of final hadronic states. The fourth column gives the CKM dependence of the amplitude A_f, using the notation of Eqs. (124), (126), and (128), with the dominant term first and the sub-dominant second. The suppression factor of the second term compared to the first is given in the last column. 'Loop' refers to a penguin-versus-tree suppression factor (it is mode-dependent and roughly $\mathcal{O}(0.2-0.3)$) and $\lambda = 0.22$ is the expansion parameter of Eq. (53).

$\bar{b} \to q\bar{q}\bar{q}'$	$B^0 \to f$	$B_s \to f$	CKM dependence ofA_f	Suppression
$\bar{b} \to \bar{c}c\bar{s}$	ψK_S	$\psi\phi$	$(V_{cb}^* V_{cs})T + (V_{ub}^* V_{us})P^u$	loop $\times \lambda^2$
$\bar{b} \to \bar{s}s\bar{s}$	ϕK_S	$\phi\phi$	$(V_{cb}^* V_{cs})P^c + (V_{ub}^* V_{us})P^u$	λ^2
$\bar{b} \to \bar{u}u\bar{s}$	$\pi^0 K_S$	K^+K^-	$(V_{cb}^* V_{cs})P^c + (V_{ub}^* V_{us})T$	λ^2/loop
$\bar{b} \to \bar{c}c\bar{d}$	D^+D^-	ψK_S	$(V_{cb}^* V_{cd})T + (V_{tb}^* V_{td})P^t$	loop
$\bar{b} \to \bar{s}s\bar{d}$	$\phi\pi$	ϕK_S	$(V_{tb}^* V_{td})P^t + (V_{cb}^* V_{cd})P^c$	$\lesssim 1$
$\bar{b} \to \bar{u}u\bar{d}$	$\pi^+\pi^-$	$\pi^0 K_S$	$(V_{ub}^* V_{ud})T + (V_{tb}^* V_{td})P^t$	loop

8 $b \to c\bar{c}s$ transitions

For $B \to J/\psi K_S$ and other $\bar{b} \to \bar{c}c\bar{s}$ processes, we can neglect the P^u contribution to $A_{\psi K}$, in the SM, to an approximation that is better than one per cent:

$$\lambda_{\psi K_S} = -e^{-2i\beta} \;\Rightarrow\; S_{\psi K_S} = \sin 2\beta\,, \quad C_{\psi K_S} = 0\,. \tag{130}$$

(Below the per cent level, several effects modify this equation [75,76].) The experimental measurements give the following ranges [77]:

$$S_{\psi K_S} = 0.69 \pm 0.03\,, \quad C_{\psi K_S} = 0.02 \pm 0.05\,. \tag{131}$$

The consistency of the experimental results (131) with the SM predictions means that the KM mechanism of CP violation has successfully passed its first precision test. For the first time, we can make the following statement based on experimental evidence:

> *Very likely, the Kobayashi–Maskawa mechanism is the dominant source of CP violation in flavour-changing processes.*

There are three qualifications implicit in this statement, and we now explain them in a little more detail [78].

- *'Very likely'*. It could be that the success is accidental. It could happen, for example, that $\sin 2\beta$ is significantly different from the SM value and that, at the same time, there is a significant CP-violating contribution to the B^0–$\overline{B}^0$ mixing amplitude, and the sum of $M_{12}^{\rm SM} + M_{12}^{\rm NP}$ accidentally carries the same phase as the one predicted by the SM alone. It could also happen that the size of NP contributions to $b \to d$ transitions is small, or that its phase is similar to the SM one, but that in $b \to s$ transitions the deviation is significant.
- *'Dominant'*. While $S_{\psi K}$ is measured with an accuracy of order 0.04, the accuracy of the SM prediction for $\sin 2\beta$ is only at the level of 0.2. Thus, it is quite possible that there is a new physics contribution at the level of $|M_{12}^{\rm NP}/M_{12}^{\rm SM}| \lesssim \mathcal{O}(0.2)$.
- *'Flavour changing'*. It may well happen that the KM phase, which is closely related to flavour violation through the CKM matrix, dominates meson decays while new, flavour-diagonal phases (such as the two unavoidable phases in the universal version of the MSSM) dominate observables such as electric dipole moments by many orders of magnitude.

The measurement of $S_{\psi K}$ provides a significant constraint on the unitarity triangle. In the ρ–η plane, it reads:

$$\sin 2\beta = \frac{2\eta(1-\rho)}{\eta^2 + (1-\rho)^2} = 0.69 \pm 0.03\ . \tag{132}$$

One can get an impression of the impact of this constraint by looking at Fig. 2, where the blue region represents $\sin 2\beta = 0.69 \pm 0.03$. An impression of the KM test can be achieved by observing that the blue region has an excellent overlap with the region allowed by all other measurements. A comparison between the constraints in the ρ–η plane from CP-conserving and CP-violating processes is provided in Fig. 4. The impressive consistency between the two allowed regions is the basis for our statement that the KM mechanism has passed its first precision tests. The fact that the allowed region from the CP-violating processes is more strongly constrained is related to the fact that CP is a good symmetry of the strong interactions and that, therefore, various CP-violating observables — in particular $S_{\psi K}$ — can be cleanly interpreted.

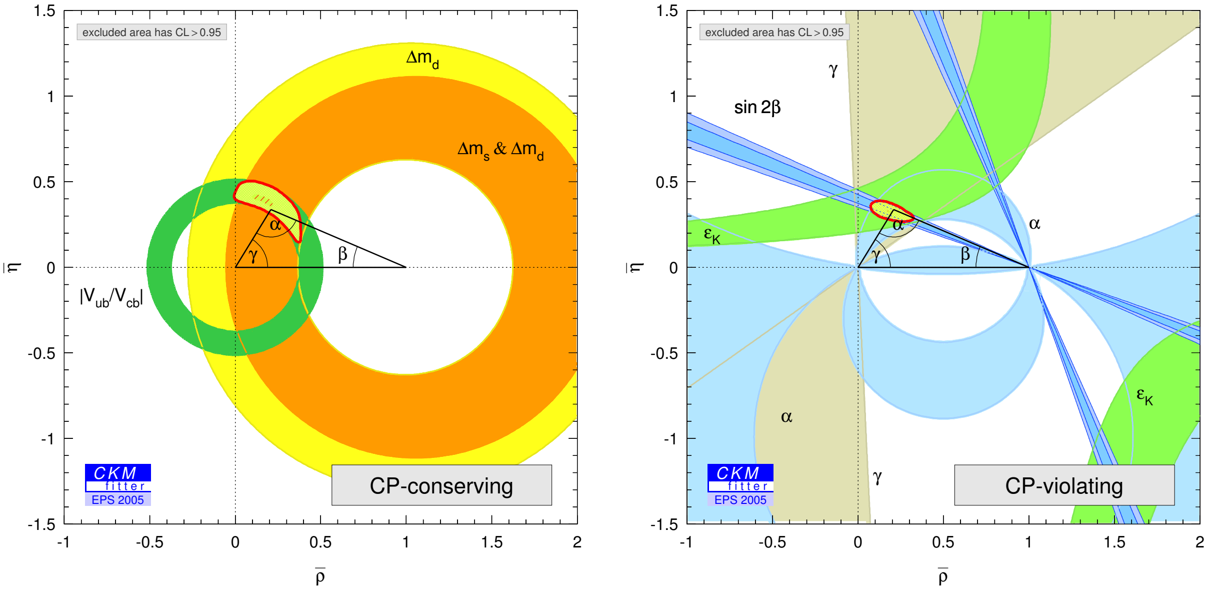


Fig. 4: Constraints in the ρ–η plane from (a) CP-conserving and (b) CP-violating loop processes

9 Penguin-dominated $b \longrightarrow s$ transitions

9.1 General considerations

The present experimental situation concerning CP asymmetries in decays to final CP eigenstates dominated by $b \longrightarrow s$ penguins is summarized in Table 2.

For $B \longrightarrow \phi K_S$ and other $\bar{b} \longrightarrow \bar{s}s\bar{s}$ processes, we can neglect the P^u contribution to A_f, in the SM, to an approximation that is good to the order of a few per cent:

$$\lambda_{\phi K_S} \approx -e^{-2i\beta} \ \Rightarrow\ S_{\phi K_S} \approx \sin 2\beta\ , \quad C_{\phi K_S} \approx 0\ . \tag{133}$$

In the presence of new physics, both A_f and M_{12} can get contributions that are comparable in size to those of the SM and carry new weak phases [36]. Such a situation gives several interesting consequences for $\bar{b} \rightarrow \bar{s}s\bar{s}$ decays.

1. A new CP-violating phase in the $b \longrightarrow s$ decay amplitude will lead to a deviation of $-\eta_f S_f$ from $S_{\psi K}$.

Table 2: CP asymmetries in $b \to s$ penguin-dominated modes

$f_{\rm CP}$	$-\eta_{f_{\rm CP}} S_{f_{\rm CP}}$	$C_{f_{\rm CP}}$
ϕK_S	$+0.47 \pm 0.19$	-0.09 ± 0.15
$\eta' K_S$	$+0.50 \pm 0.09(0.13)$	$-0.07 \pm 0.07(0.10)$
$f_0 K_S$	$+0.75 \pm 0.24$	$+0.06 \pm 0.21(0.23)$
$\pi^0 K_S$	$+0.31 \pm 0.26$	-0.02 ± 0.13
ωK_S	$+0.63 \pm 0.30$	-0.44 ± 0.24
$K_S K_S K_S$	$+0.61 \pm 0.23$	$-0.31 \pm 0.17(0.20)$

2. The S_f's will be different, in general, among the various f's. Only if the new physics contribution to A_f dominates over the SM should we expect a universal S_f.
3. A new CP-violating phase in the $b \to s$ decay amplitude in combination with a strong phase will lead to $C_f \neq 0$.

9.2 Calculating the deviations from $S_f = S_{\psi K}$

It is important to understand how large a deviation from the approximate equalities in Eq. (133) is expected within the SM. The SM contribution to the decay amplitudes, related to $\bar{b} \to \bar{q} q \bar{s}$ transitions, can always be written as a sum of two terms, $A_f^{\rm SM} = A_f^c + A_f^u$, with $A_f^c \propto V_{cb}^* V_{cs}$ and $A_f^u \propto V_{ub}^* V_{us}$. Defining the ratio $a_f^u \equiv e^{-i\gamma}(A_f^u/A_f^c)$, we have

$$A_f^{\rm SM} = A_f^c (1 + a_f^u e^{i\gamma}) \; . \tag{134}$$

The size of the deviations from Eq. (133) is set by a_f^u. For $|a_f^u| \ll 1$, we obtain

$$\begin{aligned} -\eta_f S_f &\simeq \sin 2\beta + 2\cos 2\beta\, \mathcal{R}e(a_f^u) \sin\gamma \; , \\ C_f &\simeq -2\mathcal{I}m(a_f^u) \sin\gamma . \end{aligned} \tag{135}$$

For charmless modes, the effects of the a_f^u terms (often called 'the SM pollution') are at least of order $|(V_{ub}^* V_{us})/(V_{cb}^* V_{cs})| \sim$ a few per cent.

To calculate them explicitly, we use the operator product expansion (OPE). We follow the notations of Ref. [79]. We consider the following effective Hamiltonian for $\Delta B = \pm 1$ decays:

$$\mathcal{H}_{\rm eff} = \frac{G_F}{\sqrt{2}} \sum_{p=u,c} V_{ps}^* V_{pb} \left(C_1 O_1^p + C_2 O_2^p + \sum_{i=3}^{10} C_i O_i + C_{7\gamma} O_{7\gamma} + C_{8g} O_{8g} \right) + \text{h.c.} \; , \tag{136}$$

with

$$\begin{aligned} O_1^p &= (\bar{p} b)_{V-A} (\bar{s} p)_{V-A} \; , & O_2^p &= (\bar{p}_\beta b_\alpha)_{V-A} (\bar{s}_\alpha p_\beta)_{V-A} \; , \\ O_3 &= (\bar{s} b)_{V-A} \sum_q (\bar{q} q)_{V-A} \; , & O_4 &= (\bar{s}_\alpha b_\beta)_{V-A} \sum_q (\bar{q}_\beta q_\alpha)_{V-A} \; , \\ O_5 &= (\bar{s} b)_{V-A} \sum_q (\bar{q} q)_{V+A} \; , & O_6 &= (\bar{s}_\alpha b_\beta)_{V-A} \sum_q (\bar{q}_\beta q_\alpha)_{V+A} \; , \\ O_7 &= \frac{3}{2} (\bar{s} b)_{V-A} \sum_q e_q (\bar{q} q)_{V+A} \; , & O_8 &= \frac{3}{2} (\bar{s}_\alpha b_\beta)_{V-A} \sum_q e_q (\bar{q}_\beta q_\alpha)_{V+A} \; , \end{aligned}$$

$$O_9=\frac{3}{2}(\bar sb)_{V-A}\sum_q e_q(\bar qq)_{V-A}\ ,\qquad O_{10}=\frac{3}{2}(\bar s_\alpha b_\beta)_{V-A}\sum_q e_q(\bar q_\beta q_\alpha)_{V-A}\ ,$$
$$O_{7\gamma}=-\frac{em_b}{8\pi^2}\bar s\sigma^{\mu\nu}(1+\gamma_5)F_{\mu\nu}b\ ,\qquad O_{8g}=-\frac{g_sm_b}{8\pi^2}\bar s\sigma^{\mu\nu}(1+\gamma_5)G_{\mu\nu}b\ ,\tag{137}$$

where $(\bar q_1q_2)_{V\pm A}=\bar q_1\gamma_\mu(1\pm\gamma_5)q_2$, the sum is over active quarks, with e_q denoting their electric charge in fractions of $|e|$, and α,β are colour indices. The decay amplitudes can be calculated from this effective Hamiltonian:

$$A_f=\langle f|\mathcal{H}_{\rm eff}|B^0\rangle\ ,\quad \overline{A}_f=\langle f|\mathcal{H}_{\rm eff}|\overline{B}^0\rangle\ .\tag{138}$$

The electroweak model determines the Wilson coefficients while QCD (or, more practically, a calculational method such as QCD factorization) determines the matrix elements $\langle f|O_i|B^0(\overline{B}^0)\rangle$.

Take, for example, the $B^0\to K^0\pi^0$ decay amplitude. It can be written as follows (for simplicity, we omit the contributions from O_{7-10}):

$$\begin{aligned}A^c_{K^0\pi^0}&\approx iV_{cb}^*V_{cs}\frac{G_F}{2}f_KF^{B\to\pi}(m_K^2)(m_B^2-m_\pi^2)\left(a_4+r_\chi a_6\right)\ ,\qquad(139)\\ A^u_{K^0\pi^0}&\approx iV_{ub}^*V_{us}\frac{G_F}{2}\left[f_KF^{B\to\pi}(m_K^2)(m_B^2-m_\pi^2)\left(a_4+r_\chi a_6\right)-f_\pi F^{B\to K}(m_\pi^2)(m_B^2-m_K^2)a_2\right]\ ,\end{aligned}$$

where $r_\chi=2m_K^2/[m_b(m_s+m_d)]$. The a_i parameters are related to the Wilson coefficients as follows:

$$a_i\equiv C_i+\frac{1}{N_c}C_{i\pm1}\ \text{for}\ i=\text{even, odd}\ .\tag{140}$$

Within the SM, at leading order,

$$C_1(m_W)=1\ ,\quad C_{i\neq1}(m_W)=0\ .\tag{141}$$

(Strictly speaking, $C_{7\gamma}(m_W)$ and $C_{8g}(m_W)$ are also different from zero. Their contributions to the decay processes of interest occur, however, at next-to-leading order, which we neglect here for simplicity.) To run the Wilson coefficients from the weak scale m_W to the low scale of order m_b, we use

$$\vec C(\mu)=[\alpha_s(m_W)/\alpha_s(\mu)]^{\gamma/2\beta_0}\ ,\tag{142}$$

where $\beta_0=(33-2f)/3$, with $f=5$ for $m_b\leq\mu\leq m_W$, and γ is the 12-dimensional leading-log anomalous dimension matrix which can be found, for example, in Ref. [80]. The bottom line is the following set of values for the relevant a_i parameters at the scale $\mu=m_b$:

$$a_1=1.028,\ \ a_2=0.105,\ \ a_4=-0.0233,\ \ a_6=-0.0314\ .\tag{143}$$

We use the following values for the relevant hadronic parameters:

$$f_\pi=131\,\text{MeV},\ f_K=160\,\text{MeV},\ F^{B\to\pi}(0)=0.28\ ,\quad F^{B\to K}(0)=0.34,\ \ r_\chi(m_b)=1.170\ .\tag{144}$$

Thus we can estimate $a^u_{\pi K}$:

$$a^u_{\pi K}\approx\lambda^2R_u\left(1-\frac{f_\pi}{f_K}\frac{F^{B\to K}}{F^{B\to\pi}}\frac{a_2}{a_4+r_\chi a_6}\right)\approx2.75\lambda^2R_u\approx0.052\ .\tag{145}$$

We learn that the SM and factorization predict that $-S_{\pi^0K_S}-S_{\psi K_S}\approx+0.05$.

In Table 3 we give the values of the a^u_f parameter (obtained in Ref. [80] by using factorization [79, 81, 82]) for all relevant modes.

An examination of Table 3 shows that the SM pollution is small (that is, at the naively expected level of $|(V_{ub}V_{us}^*)/(V_{cb}V_{cs}^*)|\sim$ a few per cent) for $f=\phi K_S,\ \eta'K_S$ and π^0K_S. It is larger for

Table 3: The a_f^u parameters, calculated in QCD factorization at leading log and to zeroth order in Λ/m_b (except for chirally enhanced corrections), and the SM values of S_f for $\mu = m_b$ and in parentheses the respective values for $\mu = 2m_b$ (first) and $\mu = m_b/2$ (second) if different from the central one. In the last column, the results of Ref. [83], using QCD factorization at next-to-leading order, are given. Taken from Ref. [80].

f	a_f^u [80]	$-\eta_{\rm CP} S_f$ [80]	$-\eta_{\rm CP} S_f$ [83]
ψK_S	$0.$	0.69	0.69
ϕK_S	0.019	0.71	0.71 ± 0.01
$\pi^0 K_S$	$0.052\ [0.094, 0.021]$	$0.75\ [0.79, 0.72]$	$0.76^{+0.05}_{-0.04}$
ηK_S	$0.08\ [0.16, 0.02]$	$0.78\ [0.84, 0.72]$	$0.79^{+0.11}_{-0.07}$
$\eta' K_S$	$0.007\ [-0.006, 0.019]$	$0.70\ [0.68, 0.71]$	0.70 ± 0.01
ωK_S	$0.22\ [0.37, 0.04]$	$0.88\ [0.94, 0.74]$	0.82 ± 0.08
$\rho^0 K_S$	$-0.16\ [-0.32, 0.005]$	$0.45\ [0.15, 0.70]$	$0.61^{+0.08}_{-0.12}$

$f = \eta K_S$, ωK_S and $\rho^0 K_S$. In these modes, a_f^u is enhanced because, within the QCD factorization approach, there is an accidental cancellation between the leading contributions to A_f^c. The reason for the suppression of the leading A_f^c piece in $f = \rho K$, ωK versus $f = \pi^0 K$ is that the dominant QCD-penguin coefficients a_4 and a_6 appear in $A^c_{(\rho,\omega)K}$ as $(a_4 - r_\chi a_6)$ and in $A^c_{\pi^0 K}$ as $(a_4 + r_\chi a_6)$. Since $r_\chi \simeq 1$ and, within the SM, $a_4 \sim a_6$, there is a cancellation in $A^c_{(\rho,\omega)K}$ while there is not one in $A^c_{\pi^0 K}$. The suppression for $A^c_{\eta K}$ with respect to $A^c_{\eta' K}$ has a different reason: it is due to the octet–singlet mixing, which causes destructive (constructive) interference in the $\eta(\eta')K$ penguin amplitude [84].

10 $b \to u\bar{u}d$ transitions

The present experimental situation concerning CP asymmetries in decays to final CP eigenstates via $b \to d$ transitions is summarized in Table 4.

Table 4: CP asymmetries in $b \to c\bar{c}d$ (above line) or $b \to u\bar{u}d$ (below line) modes

$f_{\rm CP}$	$-\eta_{f_{\rm CP}} S_{f_{\rm CP}}$	$C_{f_{\rm CP}}$
$\psi\pi^0$	$+0.69 \pm 0.25$	-0.11 ± 0.20
D^+D^-	$+0.29 \pm 0.63$	$+0.11 \pm 0.35$
$D^{*+}D^{*-}$	$+0.75 \pm 0.23$	-0.04 ± 0.14
$\pi^+\pi^-$	$+0.50 \pm 0.12(0.18)$	$-0.37 \pm 0.10(0.23)$
$\pi^0\pi^0$		-0.28 ± 0.39
$\rho^+\rho^-$	$+0.22 \pm 0.22$	-0.02 ± 0.17

For $B \to \pi\pi$ and other $\bar{b} \to \bar{u}u\bar{d}$ processes, the penguin-to-tree ratio can be estimated using $SU(3)$ relations and experimental data on related $B \to K\pi$ decays. The result is that the suppression is of order $0.2 - 0.3$ and so cannot be neglected. The expressions for $S_{\pi\pi}$ and $C_{\pi\pi}$ to leading order in

$R_{PT} \equiv (|V_{tb}V_{td}|P^t_{\pi\pi})/(|V_{ub}V_{ud}|T_{\pi\pi})$ are

$$\begin{aligned} \lambda_{\pi\pi} &= e^{2i\alpha}\left[(1-R_{PT}e^{-i\alpha})/(1-R_{PT}e^{+i\alpha})\right] \Rightarrow \\ S_{\pi\pi} &\approx \sin 2\alpha + 2\,\mathcal{R}e(R_{PT})\cos 2\alpha \sin\alpha\ , \quad C_{\pi\pi} \approx 2\,\mathcal{I}m(R_{PT})\sin\alpha\ . \end{aligned} \tag{146}$$

Note that R_{PT} is mode dependent and, in particular, could be different for $\pi^+\pi^-$ and $\pi^0\pi^0$. If strong phases can be neglected then R_{PT} is real, resulting in $C_{\pi\pi} = 0$. The size of $C_{\pi\pi}$ is an indicator of how large the strong phase is. With regard to $S_{\pi\pi}$, it is clear from Eq. (146) that the relative size and strong phase of the penguin contribution must be known to extract α. This is the problem of penguin pollution.

The cleanest solution involves isospin relations among the $B \to \pi\pi$ amplitudes. Let us derive this relation step by step. The $SU(2)$-isospin representations of the $\pi\pi$ states are as follows:

$$\begin{aligned} \langle\pi^+\pi^-| &= \sqrt{\frac{1}{2}}\langle(1,+1)(1,-1)+(1,-1)(1,+1)| = \sqrt{\frac{1}{3}}\,\langle 2,0| + \sqrt{\frac{2}{3}}\,\langle 0,0|\ , \\ \langle\pi^0\pi^0| &= \langle(1,0)(1,0)| = \sqrt{\frac{2}{3}}\,\langle 2,0| - \sqrt{\frac{1}{3}}\,\langle 0,0|\ , \\ \langle\pi^+\pi^0| &= \sqrt{\frac{1}{2}}\langle(1,+1)(1,0)+(1,0)(1,+1)| = \langle 2,+1|. \end{aligned} \tag{147}$$

The Hamiltonian, with its four quark operators, has two features that are important for our purposes:

1. There are $\Delta I = 1/2$ and $\Delta I = 3/2$ pieces, but no $\Delta I = 5/2$ one. The absence of the latter gives isospin relations among the $B \to \pi\pi$ amplitudes.
2. The penguin operators are purely $\Delta I = 1/2$. Thus we will find that they do not contribute to the $\pi^\pm\pi^0$ modes.

We contract the Hamiltonian with the $(B^+, B^0) = (1/2, \pm 1/2)$ states:

$$\begin{aligned} H_{3/2,+1/2}|1/2,-1/2\rangle &\propto \sqrt{\frac{1}{2}}\,|2,0\rangle + \sqrt{\frac{1}{2}}\,|1,0\rangle\ , \\ H_{3/2,+1/2}|1/2,+1/2\rangle &\propto \sqrt{\frac{3}{4}}\,|2,1\rangle - \sqrt{\frac{1}{4}}\,|1,1\rangle\ , \\ H_{1/2,+1/2}|1/2,-1/2\rangle &\propto \sqrt{\frac{1}{2}}\,|1,0\rangle - \sqrt{\frac{1}{2}}\,|0,0\rangle\ , \\ H_{1/2,+1/2}|1/2,+1/2\rangle &\propto |1,0\rangle\ . \end{aligned} \tag{148}$$

Combining Eqs. (147) and (148), we obtain

$$\begin{aligned} A_{\pi^+\pi^-} &= \sqrt{1/6}\,A_{3/2} - \sqrt{1/3}\,A_{1/2}\ , \\ A_{\pi^0\pi^0} &= \sqrt{1/3}\,A_{3/2} + \sqrt{1/6}\,A_{1/2}\ , \\ A_{\pi^+\pi^0} &= \sqrt{3/4}\,A_{3/2}\ . \end{aligned} \tag{149}$$

Analogous relations hold for the CP-conjugate amplitudes, $\overline{A}_{\pi^i\pi^j}$. These isospin decompositions lead to the Gronau–London triangle relations [85]:

$$\begin{aligned} \frac{1}{\sqrt{2}}A_{\pi^+\pi^-} + A_{\pi^0\pi^0} &= A_{\pi^+\pi^0}\ , \\ \frac{1}{\sqrt{2}}\overline{A}_{\pi^+\pi^-} + \overline{A}_{\pi^0\pi^0} &= \overline{A}_{\pi^-\pi^0}\ . \end{aligned} \tag{150}$$

The method further exploits the fact that the penguin contribution to $P_{\pi\pi}$ is pure $\Delta I = \frac{1}{2}$ (this is not true for the electroweak penguins which, however, are expected to be small), while the tree contribution to

$T_{\pi\pi}$ contains pieces which are both $\Delta I = \frac{1}{2}$ and $\Delta I = \frac{3}{2}$. A simple geometric construction then allows one to find R_{PT} and extract α cleanly from $S_{\pi^+\pi^-}$. Explicitly, one notes that, since $A_{3/2}$ comes purely from tree contributions, we have

$$\frac{q}{p}\frac{\overline{A}_{3/2}}{A_{3/2}} = -e^{2i\alpha} \; . \tag{151}$$

The branching ratios of the various modes determine $|A_{\pi^i\pi^j}|$ and $|\overline{A}_{\pi^i\pi^j}|$ (with $|A_{\pi^+\pi^0}| = |\overline{A}_{\pi^-\pi^0}|$). This would determine the shape of each of the triangles (150). Defining

$$A_0 \equiv (1/\sqrt{6})\; A_{1/2} \; , \qquad A_2 \equiv (1/\sqrt{12})\; A_{3/2} \; , \tag{152}$$

we can obtain $A_2 = (1/3)A_{\pi^+\pi^0}$ and $A_0 = (1/\sqrt{2})A_{\pi^+\pi^-} - A_2$. Similarly, we can obtain $\overline{A}_2$ and $\overline{A}_0$. Next, we define (and obtain)

$$\theta \equiv \arg(A_0 A_2^*) \; , \qquad \overline{\theta} \equiv \arg(\overline{A}_0 \overline{A}_2^*) \; . \tag{153}$$

Then we have

$$\mathcal{I}m\lambda_{\pi^+\pi^-} = \mathcal{I}m\left(-e^{-2i\alpha}\frac{|\overline{A}_2| - |\overline{A}_0|e^{i\overline{\theta}}}{|A_2| - |A_0|e^{i\theta}}\right) \; . \tag{154}$$

On the other hand, we can use the experimentally measured quantities to extract $\mathcal{I}m\lambda_{\pi^+\pi^-}$:

$$\mathcal{I}m\lambda_{\pi^+\pi^-} = \frac{S_{\pi^+\pi^-}}{1 + C_{\pi^+\pi^-}} \; . \tag{155}$$

The key experimental difficulty is that one must measure accurately the separate rates for $B^0, \overline{B}^0 \to \pi^0\pi^0$. It has been noted that an upper bound on the average rate allows one to put a useful upper bound on the deviation of $S_{\pi^+\pi^-}$ from $\sin 2\alpha$ [86–88]. Parametrizing the asymmetry by $S_{\pi^+\pi^-}/\sqrt{1-(C_{\pi^+\pi^-})^2} = \sin(2\alpha_{\text{eff}})$, the bound reads

$$\cos(2\alpha_{\text{eff}} - 2\alpha) \geq \frac{1}{\sqrt{1-(C_{\pi^+\pi^-})^2}}\left[1 - \frac{2\mathcal{B}_{00}}{\mathcal{B}_{+0}} + \frac{(\mathcal{B}_{+-} - 2\mathcal{B}_{+0} + 2\mathcal{B}_{00})^2}{4\mathcal{B}_{+-}\mathcal{B}_{+0}}\right] \; . \tag{156}$$

$\mathcal{B}_{ij}$ are averages over CP-conjugate branching ratios, e.g., $\mathcal{B}_{00} = \frac{1}{2}\left[\mathcal{B}(B^0 \to \pi^0\pi^0) + \mathcal{B}(\overline{B}^0 \to \pi^0\pi^0)\right]$. CP asymmetries in $B \to \rho\pi$ and, in particular, in $B \to \rho\rho$ can also be used to determine α [89–93]. At present, the constraints read [54]

$$\begin{aligned} |\alpha_{\text{eff}}^{\pi^+\pi^-} - \alpha| &< 38^\circ \; , & R_{PT}^{\pi^+\pi^-} &= 0.37 \pm 0.17 \; , \\ |\alpha_{\text{eff}}^{\rho^+\rho^-} - \alpha| &< 14^\circ \; , & R_{PT}^{\rho^+\rho^-} &= 0.07^{+0.14}_{-0.07} \; . \end{aligned} \tag{157}$$

Using isospin analyses for all three systems ($\pi\pi$, $\rho\pi$ and $\rho\rho$), one obtains [54]

$$\alpha(\pi\pi, \pi\rho, \rho\rho) = \left[101^{+16}_{-9}\right]^\circ \; , \tag{158}$$

to be compared with the result of the CKM fit,

$$\alpha(\text{CKM fit}) = 96 \pm 16^\circ \; . \tag{159}$$

We would like to emphasize the following points:

- The consistency of Eq. (158) with Eq. (159) means that *the KM mechanism of CP violation has successfully passed a second precision test.*

- The α measurement via the $b \to u\bar{u}d$ transitions provides a significant constraint on the unitarity triangle.
- The isospin analysis determines the relative phase between the B^0–$\overline{B}^0$ mixing amplitude and the tree decay amplitude $A_{3/2}$, independent of the electroweak model. The tree decay amplitude is not affected by new physics. Any new physics modification of the mixing amplitude is measured by $S_{\psi k}$. Thus, the combination of $S_{\psi K}$ and the isospin analysis of $S_{\pi\pi,\rho\pi,\rho\rho}$ constrains α even in the presence of new physics in B^0–$\overline{B}^0$ mixing.

11 $b \to c\bar{u}s, u\bar{c}s$ transitions

An interesting set of measurements is that of $B \to DK$ which proceed via the quark transitions $\bar{b} \to \bar{c}u\bar{s}$ or $\bar{b} \to \bar{u}c\bar{s}$ (and their CP conjugates). Given the quark processes, it is clear that there is no penguin contribution here. Thus, the quark transitions are purely tree processes. The interference between the two quark transitions (if they lead to the same final states, see below) is sensitive to $\arg[(V_{ub}^* V_{us})/(V_{cb}^* V_{cs})] \approx \gamma$.

There are three variants on this method: GLW [94, 95], ADS [96] and GGSZ [97]. The simplest one to explain involves branching ratios of charged B decays, and thus B^0–$\overline{B}^0$ mixing plays no role. Consider the decay $B^\pm \to D_1^0 K^\pm$, where $D_{1,2}^0 = \frac{1}{\sqrt{2}}(D^0 \pm \overline{D}^0)$ are the CP eigenstates. Taking into account that

$$\begin{aligned} A(B^+ \to D^0 K^+) \times A(D^0 \to D_1^0) &\propto (V_{ub}^* V_{cs}) \times (V_{cs}^* V_{us}) \, , \\ A(B^+ \to \overline{D}^0 K^+) \times A(\overline{D}^0 \to D_1^0) &\propto (V_{cb}^* V_{us}) \times (V_{us}^* V_{cs}) \, , \end{aligned} \tag{160}$$

we can write the relevant decay amplitudes as follows:

$$\begin{aligned} \sqrt{2} A_{D_1^0 K^+} &= |A_{D^0 K^+}| e^{i(\delta+\gamma)} + |A_{\overline{D}^0 K^+}| = A_{D^0 K^+} + A_{\overline{D}^0 K^+} \, , \\ \sqrt{2} A_{D_1^0 K^-} &= |A_{D^0 K^-}| e^{i(\delta-\gamma)} + |A_{\overline{D}^0 K^-}| = A_{\overline{D}^0 K^-} + A_{D^0 K^-} \, . \end{aligned} \tag{161}$$

Measuring the rates for the six relevant decay modes ($D_1^0 K^+$, $D^0 K^+$, $\overline{D}^0 K^+$ and the CP-conjugate modes), one can construct an amplitude triangle for each of the two relations in Eq. (161). We can choose a phase convention where $A_{\overline{D}^0 K^+} = A_{D^0 K^-}$. Then, the relative angle between $A_{D^0 K^+}$ and $A_{\overline{D}^0 K^-}$ is 2γ.

The method of Ref. [97] gives, at present, the most significant constraints. It allows one to determine the amplitude ratios, $r_B = 0.12^{+0.03}_{-0.04}$ and $r_B^* = 0.09^{+0.03}_{-0.04}$, and the weak phase γ [54]:

$$\gamma(DK) = (63^{+15}_{-13})^\circ \, . \tag{162}$$

This range is to be compared with the range of γ derived from the CKM fit (not including the direct γ measurements):

$$\gamma(\text{CKM fit}) = (57^{+7}_{-14})^\circ \, . \tag{163}$$

We would like to emphasize the following points:

- The consistency of Eq. (162) with Eq. (163) means that *the KM mechanism of CP violation has successfully passed a third precision test.*
- The γ measurement via the $b \to c\bar{u}s, u\bar{c}s$ y transitions provides yet another constraint on the unitarity triangle. The constraint will become more significant when the experimental precision improves.
- The determination of γ here relies on tree decay amplitudes. Thus, the analysis of $B \to DK$ decays constrains γ even in the presence of new physics in loop processes.

12 CP violation as a probe of new physics

We have argued that the SM picture of CP violation is unique and highly predictive. We have also stated that reasonable extensions of the SM have a very different picture of CP violation. Experimental results are now starting to decide between the various possibilities. Our discussion of CP violation in the presence of new physics is aimed to demonstrate that, indeed, models of new physics can significantly modify the SM predictions and that present and near future measurements have therefore a strong impact on the theoretical understanding of CP violation.

To understand how the SM predictions could be modified by new physics, we focus on CP violation in the interference between decays with and without mixing. As explained above, this type of CP violation may give, owing to its theoretical cleanliness, unambiguous evidence for new physics most easily. We now demonstrate what type of questions can be (or have already been) answered when many such observables are measured.

I. Consider $S_{\psi K_S}$, the CP asymmetry in $B \to \psi K_S$. This measurement cleanly determines the relative phase between the B^0–$\overline{B}^0$ mixing amplitude and the $b \to c\bar{c}s$ decay amplitude ($\sin 2\beta$ in the SM). The $b \to c\bar{c}s$ decay has Standard Model tree contributions and therefore is very unlikely to be significantly affected by new physics. On the other hand, the mixing amplitude can be easily modified by new physics. We parametrize such a modification as follows:

$$r_d^2\, e^{2i\theta_d} = \frac{M_{12}}{M_{12}^{\rm SM}}\,. \tag{164}$$

Then the following observables provide constraints on r_d^2 and $2\theta_d$:

$$\begin{aligned}
S_{\psi K_S} &= \sin(2\beta + 2\theta_d)\,,\\
\Delta m_B &= r_d^2(\Delta m_B)^{\rm SM}\,,\\
\mathcal{A}_{\rm SL} &= -\mathcal{R}e\left(\frac{\Gamma_{12}}{M_{12}}\right)^{\rm SM}\frac{\sin 2\theta_d}{r_d^2} + \mathcal{I}m\left(\frac{\Gamma_{12}}{M_{12}}\right)^{\rm SM}\frac{\cos 2\theta_d}{r_d^2}\,.
\end{aligned} \tag{165}$$

Examining whether $S_{\psi K_S}$, Δm_B and $\mathcal{A}_{\rm SL}$ fit the SM prediction, that is, whether $\theta_d \neq 0$ and/or $r_d^2 \neq 1$, we can answer the following question (see, for example, Ref. [98]):

(i) *Is there new physics in B^0–$\overline{B}^0$ mixing?*

Thanks to the fact that quite a few observables that are related to SM tree-level processes have already been measured, we are able to refer to this question in a quantitative way. The tree-level processes are insensitive to new physics and can be used to constrain ρ and η even in the presence of new physics contributions to loop processes, such as Δm_B. Among these observables we have $|V_{cb}|$ and $|V_{ub}|$ from semileptonic B decays, the phase γ from $B \to DK$ decays, and the phase α from $B \to \rho\rho$ decays (in combination with $S_{\psi K}$). One can fit these observables, and the ones in Eq. (165), to the four parameters ρ, η, r_d^2 and $2\theta_d$. The resulting constraints are shown in Fig. 5.

A long list of models that require a significant modification of the B^0–$\overline{B}^0$ mixing amplitude are excluded. We can further conclude from Fig. 5 that a new physics (NP) contribution to the B^0–$\overline{B}^0$ mixing amplitude at a level higher than 20% is now disfavoured. Yet, it is still possible that ρ and η are well outside their SM range and that NP gives $2\theta_d$ very different from zero and/or r_d^2 very different from one. In this case, the SM and the NP 'conspire' to mimic the SM values of the observables (165). This is what we meant concretely in our statement that the KM dominance of the observed CP violation is now very likely but not guaranteed.

II. Consider $S_{\phi K_S}$, the CP asymmetry in $B \to \phi K_S$. This measurement is sensitive to the relative phase between the B–$\bar{B}$ mixing amplitude and the $b \to s\bar{s}s$ decay amplitude ($\sin 2\beta$ in the SM). The $b \to s\bar{s}s$ decay has only SM penguin contributions and therefore is sensitive to NP. For the simple case

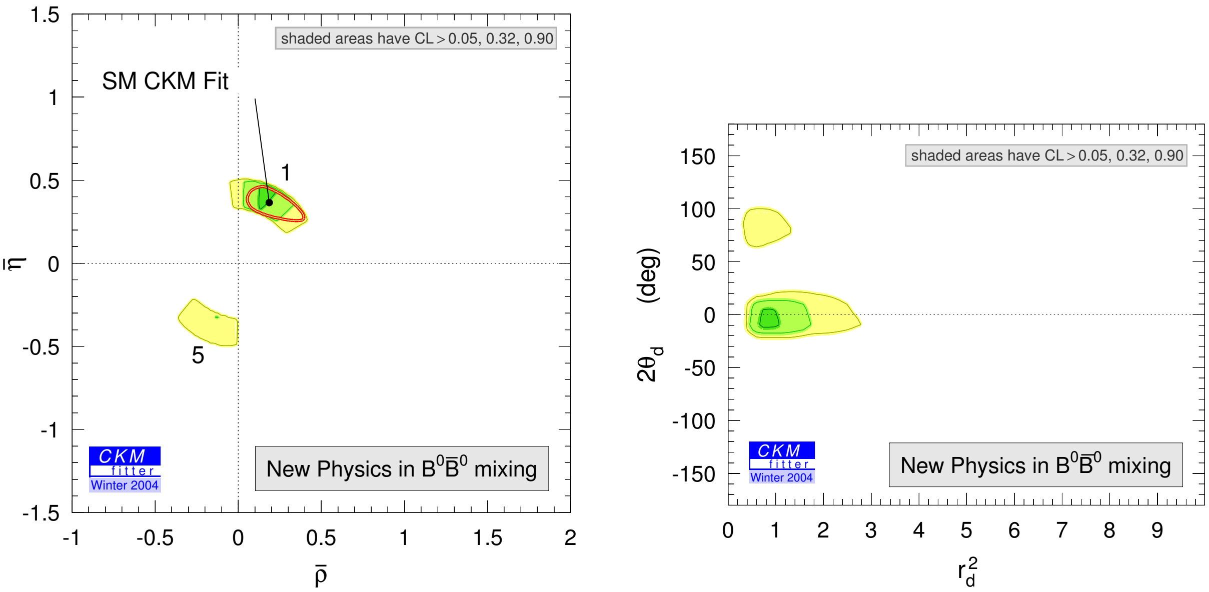


Fig. 5: Constraints in the (a) ρ–η plane and (b) r_d^2–$2\theta_d$ plane, assuming that NP contributions to tree-level processes are negligible [54]

that the NP contribution depends on a single CP-violating phase ϕ_{bs}, we parametrize the modification of the decay amplitude as follows [for simplicity, we neglect here the a_f^u terms of Eq. (134)]:

$$A_f = A_f^c \left(1 + b_f \, e^{i\phi_{bs}}\right) \; . \tag{166}$$

Here b_f is complex only if it carries a strong phase. The effects of this NP contribution are simple to understand in two limits.

1. The NP contribution is dominant, $|b_f| \gg 1$. The shift in all modes where this condition is valid is universal and depends only on ϕ_{bs}:

$$\begin{aligned} -\eta_f S_f &\simeq \sin(2\beta+2\theta_d)\cos 2\phi_{bs} + \cos(2\beta+2\theta_d)\sin 2\phi_{bs} \; , \\ C_f &\simeq 0 \; . \end{aligned} \tag{167}$$

2. The NP contribution is small. Explicitly, $|b_f| \ll 1$. The shift is mode dependent and depends on both b_f and $\sin\phi_{bs}$:

$$\begin{aligned} -\eta_f S_f &\simeq \sin(2\beta+2\theta_d) + 2\cos(2\beta+2\theta_d)\mathcal{R}e(b_f)\sin\phi_{bs} \; , \\ C_f &\simeq -2\mathcal{I}m(b_f^c)\sin\phi_{bs} \; . \end{aligned} \tag{168}$$

Note that the effect of the NP is similar to that of the SM a_f^u terms (with $b_f \leftrightarrow a_f^u$ and $\phi_{bs} \leftrightarrow \gamma$), so that the latter have to be known in order to probe the b_f terms. Once that is done, the value of $S_{\psi K}$ determines $2\beta + 2\theta_d$ and one can examine whether $\phi_{bs} \neq 0$ and answer the following question:

(ii) *Is there new physics in $b \to s$ transitions?*

So far, the experimental data — see Table 2 — do not provide any evidence for $\phi_{bs} \neq 0$. Yet, the experimental accuracy is still not sufficient to make qualitative statements such as we made for $b \to d$ transitions (B^0–$\overline{B}^0$ mixing). To see this, we compare the constraints in the ρ–η plane that arise from tree plus $b \to d$ loops (Δm_B, $S_{\psi K_S}$, $S_{\rho\rho}$, etc.) to those from tree plus $b \to s$ loops ($S_{\phi K_S}$, $S_{\eta' K_S}$, Δm_s). This is done in Fig. 6.

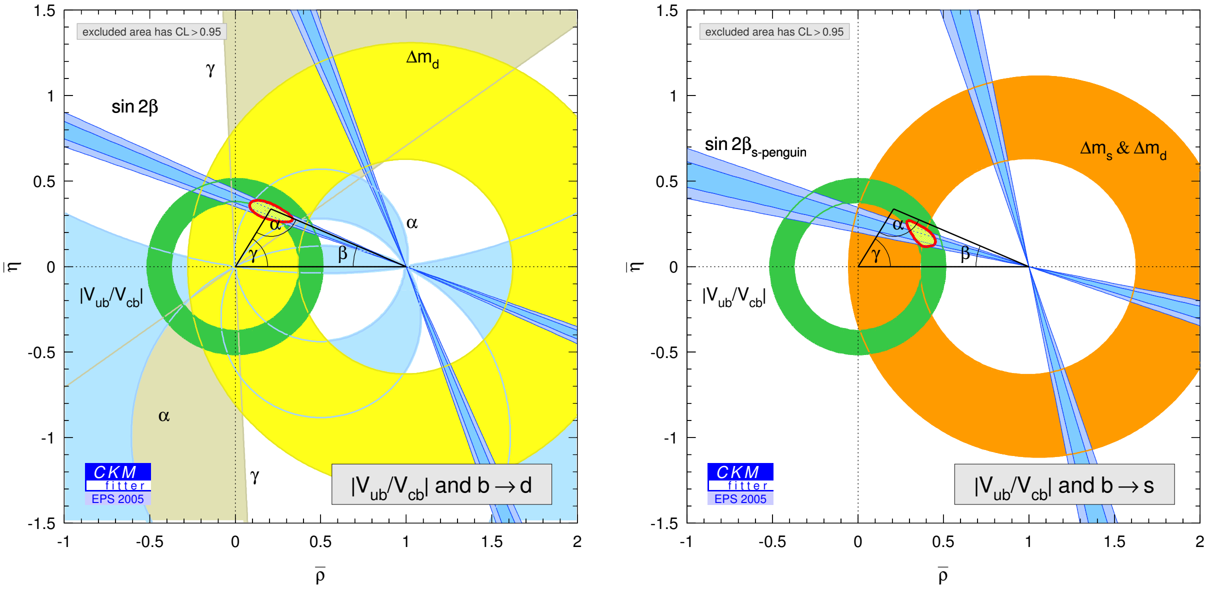


Fig. 6: Constraints in the ρ–η plane from tree processes and (a) $b \to d$ or (b) $b \to s$ loop processes

III. Together with a future measurement of B_s–$\overline{B}_s$ mixing, we may also try to answer the following question:

(iii) *Is there new physics in* $\Delta B = 1$ *processes? in* $\Delta B = 2$*? in both?*

IV. Consider $a_{\pi\nu\bar\nu} \equiv \Gamma_{K_L\to\pi^0\nu\bar\nu}/\Gamma_{K^+\to\pi^+\nu\bar\nu}$, see Eq. (111). This measurement will cleanly determine the relative phase between the K^0–$\overline{K}^0$ mixing amplitude and the $s \to d\nu\bar\nu$ decay amplitude (of order $\sin^2\beta$ in the SM). The experimentally measured small value of ε_K requires that the phase of the K^0–$\overline{K}^0$ mixing amplitude is not modified from the SM prediction. (More precisely, it requires that the phase of the mixing amplitude is very close to twice the phase of the $s \to d\bar{u}u$ decay amplitude [99].) On the other hand, the decay, which in the SM is a loop process with small mixing angles, can be easily modified by new physics. Examining whether the SM correlation between $a_{\pi\nu\bar\nu}$ and $S_{\psi K_S}$ is fulfilled, we can answer the following question:

(iv) *Is there new physics related solely to the third generation? to all generations?*

To understand the present situation, we present in Fig. 7 the constraints in the ρ–η plane from tree plus loop processes that do not involve external third generation quarks, namely $s \to d$ transitions only [ϵ and $\mathcal{B}(K^+ \to \pi^+\nu\bar\nu)$]. This can be compared with the constraints from tree plus loop processes that do involve the third generation, namely $b \to d$ and $b \to s$ transitions. Again, one can see that there is a lot to be learned from future measurements. (For a recent, comprehensive analysis of this question, see Ref. [100].)

V. Consider ϕ_D, defined in Eq. (117), which is the relative phase between the D^0–$\overline{D}^0$ mixing amplitude and the $c \to d\bar{s}u$ and $c \to s\bar{d}u$ decay amplitudes. Within the SM, the two decay channels are tree level. It is unlikely that they are affected by NP. On the other hand, the mixing amplitude can be easily modified by NP. Examining whether $\phi_D \neq 0$, we can answer the following question:

(v) *Is there new physics in the down sector? in the up sector? in both?*

VI. Consider d_N, the electric dipole moment of the neutron. We have not discussed this quantity so far because, unlike CP violation in meson decays, flavour-changing couplings are not necessary for d_N. In other words, the CP violation that induces d_N is *flavour diagonal.* It does in general get contributions from flavour-changing physics, but it could be induced by sectors that are flavour blind. Within the SM (and ignoring $\theta_{\rm QCD}$), the contribution from $\delta_{\rm KM}$ arises at the three-loop level and is at least six orders of

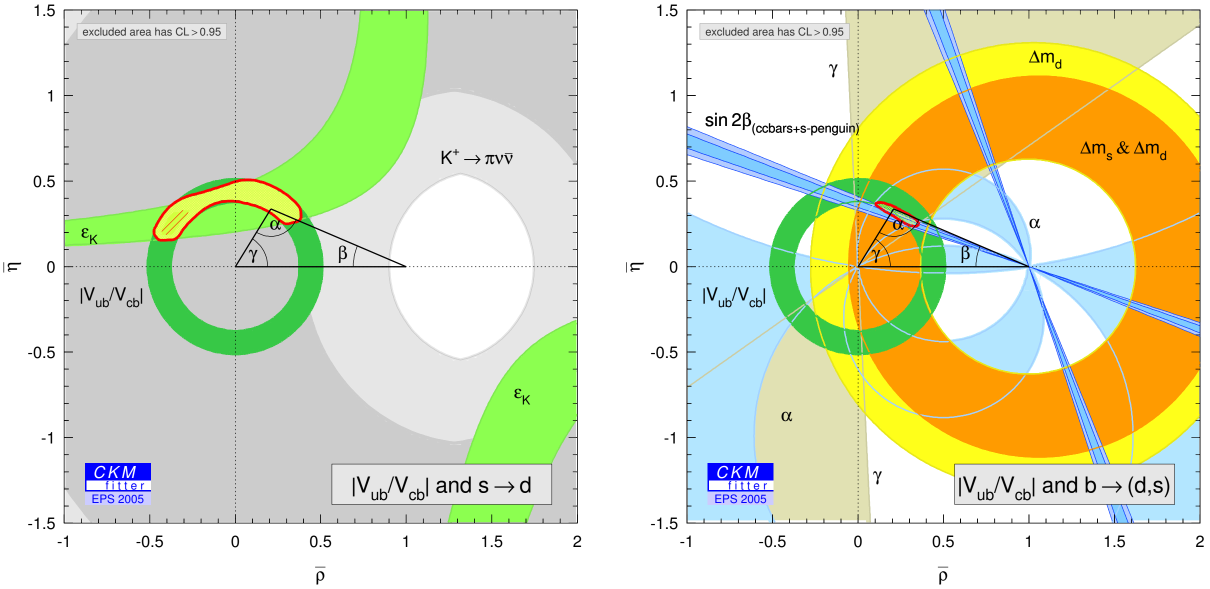


Fig. 7: Constraints in the ρ–η plane from tree processes and (a) $s \longrightarrow d$ or (b) $b \longrightarrow d$ and $b \longrightarrow s$ loop processes

magnitude below the experimental bound (13). If the bound is further improved (or a signal observed), we can answer the following question:

(vi) *Are there new sources of CP violation that are flavour changing? flavour diagonal? both?*

It is no wonder then that with such rich information, flavour and CP violation provide an excellent probe of NP. We next demonstrate this situation more concretely by discussing CP violation in supersymmetry.

13 Supersymmetric CP violation

Supersymmetry solves the fine-tuning problem of the SM and has many other virtues. But at the same time, it leads to new problems: baryon-number violation, lepton-number violation, large flavour-changing neutral-current processes and large CP violation. The first two problems can be solved by imposing R-parity on supersymmetric models. There is no such simple, symmetry-related solution to the problems of flavour and CP violation. Instead, suppression of the relevant couplings can be achieved by demanding very constrained structures of the soft supersymmetry breaking terms. There are two important questions here: First, can theories of dynamical supersymmetry breaking naturally induce such structures? Second, can measurements of flavour-changing and/or CP-violating processes shed light on the structure of the soft supersymmetry breaking terms? Since the answer to both questions is in the affirmative, we conclude that flavour-changing neutral-current processes and, in particular, CP-violating observables will provide clues to the crucial question of how supersymmetry breaks.

13.1 CP-violating parameters

A generic supersymmetric extension of the SM contains a host of new flavour- and CP-violating parameters. (For a review of CP violation in supersymmetry see Refs. [101, 102].) It is an amusing exercise to count the number of parameters [103]. The supersymmetric part of the Lagrangian depends, in addition to the three gauge couplings of $G_{\rm SM}$, on the parameters of the superpotential W:

$$W = \sum_{i,j} \left(Y^u_{ij} H_u Q_{Li} \overline{U}_{Lj} + Y^d_{ij} H_d Q_{Li} \overline{D}_{Lj} + Y^\ell_{ij} H_d L_{Li} \overline{E}_{Lj} \right) + \mu H_u H_d \, . \qquad (169)$$

In addition, we have to add soft supersymmetry breaking terms:

$$\begin{aligned}\mathcal{L}_{\text{soft}} = \; & - \; \left(A^u_{ij} H_u \tilde{Q}_{Li} \tilde{\overline{U}}_{Lj} + A^d_{ij} H_d \tilde{Q}_{Li} \tilde{\overline{D}}_{Lj} + A^\ell_{ij} H_d \tilde{L}_{Li} \tilde{\overline{E}}_{Lj} + B H_u H_d + \text{h.c.}\right) \\ & - \sum_{\text{all scalars}} (m^2_S)_{ij} A_i \bar{A}_j - \frac{1}{2} \sum_{(a)=1}^{3} \left(\tilde{m}_{(a)} (\lambda\lambda)_{(a)} + \text{h.c.}\right) \; , \end{aligned} \tag{170}$$

where $S = Q_L, \overline{D}_L, \overline{U}_L, L_L, \overline{E}_L$. The three Yukawa matrices Y^f depend on 27 real and 27 imaginary parameters. Similarly, the three A^f matrices depend on 27 real and 27 imaginary parameters. The five m^2_S Hermitian 3 × 3 mass-squared matrices for sfermions have 30 real parameters and 15 phases. The gauge and Higgs sectors depend on

$$\theta_{\text{QCD}}, \tilde{m}_{(1)}, \tilde{m}_{(2)}, \tilde{m}_{(3)}, g_1, g_2, g_3, \mu, B, m^2_{h_u}, m^2_{h_d} \; , \tag{171}$$

that is, 11 real and 5 imaginary parameters. Summing over all sectors, we get 95 real and 74 imaginary parameters. The various couplings (other than the gauge couplings) can be thought of as spurions that break a global symmetry,

$$U(3)^5 \times U(1)_{\text{PQ}} \times U(1)_R \; \to \; U(1)_B \times U(1)_L \; . \tag{172}$$

The $U(1)_{\text{PQ}} \times U(1)_R$ charge assignments are

$$\begin{array}{rccccc} & H_u & H_d & Q\overline{U} & Q\overline{D} & L\overline{E} \\ U(1)_{\text{PQ}} & 1 & 1 & -1 & -1 & -1 \\ U(1)_{\text{R}} & 1 & 1 & 1 & 1 & 1 \end{array} \; . \tag{173}$$

Consequently, we can remove 15 real and 30 imaginary parameters, which leaves

$$124 \; = \; \begin{cases} 80 & \text{real} \\ 44 & \text{imaginary} \end{cases} \quad \text{physical parameters} \; . \tag{174}$$

In particular, there are 43 new CP-violating phases! In addition to the single Kobayashi–Maskawa of the SM, we can put three phases in M_1, M_2, μ (we used the $U(1)_{\text{PQ}}$ and $U(1)_R$ to remove the phases from μB^* and M_3, respectively) and the other 40 phases appear in the mixing matrices of the fermion–sfermion–gaugino couplings. (Of the 80 real parameters, there are 11 absolute values of the parameters in (171), 9 fermion masses, 21 sfermion masses, 3 CKM angles and 36 SCKM angles.) Supersymmetry provides a nice example of our statement that reasonable extensions of the SM may have more than one source of CP violation.

The requirement of consistency with experimental data provides strong constraints on many of these parameters. For this reason, the physics of flavour and CP violation has had a profound impact on supersymmetric model building. A discussion of CP violation in this context can hardly avoid addressing the flavour problem itself. Indeed, many of the supersymmetric models that we analyse below were originally aimed at solving flavour problems.

With regard to CP violation, one can distinguish two classes of experimental constraints. First, bounds on nuclear and atomic electric dipole moments determine what is usually called the *supersymmetric CP problem*. Second, the physics of neutral mesons and, most importantly, the small experimental value of ε_K pose the *supersymmetric* ε_K *problem*. In the next two subsections we describe the two problems.

13.2 The supersymmetric CP problem

One aspect of supersymmetric CP violation involves effects that are flavour preserving. Then, for simplicity, we describe this aspect in a supersymmetric model without additional flavour mixings, i.e., the minimal supersymmetric standard model (MSSM) with universal sfermion masses and with the trilinear supersymmetry-breaking scalar couplings proportional to the corresponding Yukawa couplings. (The generalization to the case of non-universal soft terms is straightforward.) In such a constrained framework, there are four new phases beyond the two phases of the SM ($\delta_{\rm KM}$ and $\theta_{\rm QCD}$). One arises in the bilinear μ-term of the superpotential (169), while the other three arise in the soft supersymmetry breaking parameters of Eq. (170): $\tilde{m}$ (the gaugino mass), A (the trilinear scalar coupling) and B (the bilinear scalar coupling). Only two combinations of the four phases are physical [104, 105]:

$$\phi_A = \arg(A^*\tilde{m}) \ , \quad \phi_B = \arg(\tilde{m}\mu B^*) \ . \tag{175}$$

In the more general case of non-universal soft terms, there is one independent phase ϕ_{A_i} for each quark and lepton flavour. Moreover, complex off-diagonal entries in the sfermion mass-squared matrices represent additional sources of CP violation.

The most significant effect of ϕ_A and ϕ_B is their contribution to electric dipole moments (EDMs). For example, the contribution from one-loop gluino diagrams to the down-quark EDM is given by [106, 107]

$$d_d = m_d \frac{e\alpha_3}{18\pi\tilde{m}^3}\left(|A|\sin\phi_A + \tan\beta|\mu|\sin\phi_B\right) \ , \tag{176}$$

where we have taken $m^2_{\tilde{Q}} \sim m^2_{\tilde{D}} \sim m^2_{\tilde{g}} \sim \tilde{m}^2$, for left- and right-handed squark and gluino masses. We define, as usual, $\tan\beta = \langle H_u\rangle/\langle H_d\rangle$. Similar one-loop diagrams give rise to chromoelectric dipole moments. The electric and chromoelectric dipole moments of the light quarks (u, d, s) are the main source of d_N (the EDM of the neutron), giving [108]

$$d_N \sim 2\left(\frac{100 \text{ GeV}}{\tilde{m}}\right)^2 \sin\phi_{A,B} \times 10^{-23} \text{ e cm} \ , \tag{177}$$

where, as above, $\tilde{m}$ represents the overall supersymmetry (SUSY) scale. In a generic supersymmetric framework, we expect $\tilde{m} = \mathcal{O}(m_Z)$ and $\sin\phi_{A,B} = \mathcal{O}(1)$. Then the constraint (13) is generically violated by about two orders of magnitude. This is *the supersymmetric CP problem.*

Equation (177) shows two possible ways to solve the supersymmetric CP problem:

1. heavy squarks, $\tilde{m} \gtrsim 1$ TeV;
2. approximate CP, $\sin\phi_{A,B} \ll 1$.

13.3 The supersymmetric ε_K problem

The supersymmetric contribution to the ε_K parameter is dominated by diagrams involving Q and $\bar{d}$ squarks in the same loop. For $\tilde{m} = m_{\tilde{g}} \simeq m_Q \simeq m_D$ (our results depend only weakly on this assumption) and focusing on the contribution from the first two squark families, one gets (see, for example, [109])

$$\varepsilon_K = \frac{5\,\alpha_3^2}{162\sqrt{2}}\frac{f_K^2 m_K}{\tilde{m}^2\Delta m_K}\left[\left(\frac{m_K}{m_s+m_d}\right)^2 + \frac{3}{25}\right]\mathcal{I}m(\delta^d_{12})_{LL}(\delta^d_{12})_{RR}) \ . \tag{178}$$

Here

$$(\delta^d_{12})_{LL} \quad = \quad \left(\frac{m^2_{\tilde{Q}_2} - m^2_{\tilde{Q}_1}}{m^2_{\tilde{Q}}}\right) K^{dL}_{12} \ ,$$

$$(\delta^d_{12})_{RR} \;=\; \left(\frac{m^2_{\tilde{D}_2} - m^2_{\tilde{D}_1}}{m^2_{\tilde{D}}}\right) K^{dR}_{12}\;, \tag{179}$$

where K^{dL}_{12} (K^{dR}_{12}) are the mixing angles in the gluino couplings to left-handed (right-handed) down quarks and their scalar partners. Note that CP would be violated even if there were two families only [110]. Using the experimental value of ε_K, we get

$$\frac{(\Delta m_K \varepsilon_K)^{\mathrm{SUSY}}}{(\Delta m_K \varepsilon_K)^{\mathrm{EXP}}} \sim 10^7 \left(\frac{300\text{ GeV}}{\tilde{m}}\right)^2 \left(\frac{m^2_{\tilde{Q}_2} - m^2_{\tilde{Q}_1}}{m^2_{\tilde{Q}}}\right) \left(\frac{m^2_{\tilde{D}_2} - m^2_{\tilde{D}_1}}{m^2_{\tilde{D}}}\right) |K^{dL}_{12} K^{dR}_{12}| \sin\phi\;, \tag{180}$$

where ϕ is the CP-violating phase. In a generic supersymmetric framework, we expect $\tilde{m} = \mathcal{O}(m_Z)$, $\delta m^2_{Q,D}/m^2_{Q,D} = \mathcal{O}(1)$, $K^{Q,D}_{ij} = \mathcal{O}(1)$ and $\sin\phi = \mathcal{O}(1)$. Then the constraint (180) is generically violated by about seven orders of magnitude.

The Δm_K constraint on $\mathcal{R}e\big[(\delta^d_{12})_{LL}(\delta^d_{12})_{RR}\big]$ is about two orders of magnitude weaker. One can distinguish then three interesting regions for $\langle\delta^d_{12}\rangle = \sqrt{(\delta^d_{12})_{LL}(\delta^d_{12})_{RR}}$:

$$\langle\delta^d_{12}\rangle \begin{cases} \gg 0.003 & \text{excluded;} \\ \in [0.0002, 0.003] & \text{viable with small phases;} \\ \ll 0.0002 & \text{viable with } \mathcal{O}(1) \text{ phases.} \end{cases} \tag{181}$$

The first bound comes from the Δm_K constraint (assuming that the relevant phase is not particularly close to $\pi/2$). The bounds here apply to squark masses of order 500 GeV and scale like $\tilde{m}$. There is also dependence on $m_{\tilde{g}}/\tilde{m}$, which is here taken to be one.

Equation (180) also shows the possible ways to solve the supersymmetric ε_K problem:

1. heavy squarks, $\tilde{m} \gg 300$ GeV;
2. universality, $\delta m^2_{Q,D} \ll m^2_{Q,D}$;
3. alignment, $|K^d_{12}| \ll 1$;
4. approximate CP, $\sin\phi \ll 1$.

13.4 More on supersymmetric flavour and CP violation

The flavour and CP constraints on supersymmetric models apply to almost all flavour-changing couplings. The size of supersymmetric flavour violation depends on the overall scale of the soft supersymmetry breaking terms, on mass degeneracies between sfermion generations, and on the mixing angles in gaugino couplings. One can choose a representative scale (say, $\tilde{m} \sim 300$ GeV) and then conveniently present the constraints in terms of the $(\delta^q_{ij})_{MN}$ parameters, see Eq. (179). In a given supersymmetric flavour model, one can find predictions for the $(\delta^q_{ij})_{MN}$ and test the model.

A summary of upper bounds on the supersymmetric flavour-changing couplings is given in Table 5. The bounds on the $\mathcal{I}m(\delta^d_{12})_{LR,RL}$ parameters are taken from Ref. [111], on δ^d_{13} from Ref. [112], and on δ^d_{23} from Refs. [113, 114]. The bounds are expressed in powers of the Wolfenstein parameter λ, which makes it easy to compare with model predictions. As an example, we give the range of these parameters that is expected in a large class of viable models of alignment [115–117].

Until some time ago, the δ^d_{23} parameters had been only weakly constrained (the improving accuracy of the measurements of $\mathcal{B}(B \to X\ell^+\ell^-)$ has strengthened the constraints considerably). Furthermore, measurements of various CP asymmetries in penguin-dominated modes (particularly $S_{\phi K}$ and $S_{\eta' K}$) gave central values that were far off the expected value $\sim S_{\psi K}$ (at present the strongest discrepancy is down to the 2σ level). One may still ask whether effects of order 0.1, which is the order of

Table 5: Theoretical predictions for supersymmetric flavour-changing couplings in viable models of alignment, and the experimental constraints

$(\delta^q_{MN})_{ij}$	Prediction	Upper bound	$(\delta^d_{MN})_{ij}$	Prediction	Upper bound
$(\delta^d_{LL})_{12}$	$\lambda^5 - \lambda^3$	λ^3	$(\delta^d_{LR})_{12}$	$\lambda^7(m_b/\tilde{m})$	$\lambda^7(\mathcal{I}m)$
$(\delta^d_{RR})_{12}$	$\lambda^7 - \lambda^3$	$\lambda^{10}/(\delta^d_{LL})_{12}$	$(\delta^d_{RL})_{12}$	$\lambda^9(m_b/\tilde{m})$	$\lambda^7(\mathcal{I}m)$
$(\delta^d_{LL})_{13}$	λ^3	λ	$(\delta^d_{LR})_{13}$	$\lambda^3(m_b/\tilde{m})$	λ^2
$(\delta^d_{RR})_{13}$	$\lambda^7 - \lambda^3$	$\lambda^4/(\delta^d_{LL})_{13}$	$(\delta^d_{RL})_{13}$	$\lambda^7(m_b/\tilde{m})$	λ^2
$(\delta^d_{LL})_{23}$	λ^2	$\lambda^2(\mathcal{R}e) - \lambda(\mathcal{I}m)$	$(\delta^d_{LR})_{23}$	$\lambda^2(m_b/\tilde{m})$	$\lambda^4(\mathcal{R}e) - \lambda^3(\mathcal{I}m)$
$(\delta^d_{RR})_{23}$	$\lambda^4 - \lambda^2$	1	$(\delta^d_{RL})_{23}$	$\lambda^4(m_b/\tilde{m})$	λ^3
$(\delta^u_{LL})_{12}$	λ	λ			
$(\delta^u_{RR})_{12}$	$\lambda^4 - \lambda^2$	$\lambda^4/(\delta^u_{LL})_{12}$			

the expected experimental accuracy and probably above the theoretical error on $S_{\phi K}$ and $S_{\eta' K}$, are still possible within supersymmetric flavour models and, in particular, alignment models.

To answer this question, we use the results of Ref. [113]. From their Fig. 3, we make the following estimates:

$$\begin{aligned} \frac{\Delta S_{\phi K}}{\Delta \mathcal{I}m(\delta^d_{LL})_{23}} &\sim \frac{\Delta S_{\phi K}}{\Delta \mathcal{I}m(\delta^d_{RR})_{23}} \sim 0.3 \,, \\ \frac{\Delta S_{\phi K}}{\Delta \mathcal{I}m(\delta^d_{LR})_{23}} &\sim \frac{\Delta S_{\phi K}}{\Delta \mathcal{I}m(\delta^d_{RL})_{23}} \sim 100 \,. \end{aligned} \tag{182}$$

Thus, for $S_{\phi K}$ to be shifted by $\mathcal{O}(0.1)$, we need at least one of the following four options:

$$\begin{aligned} \mathcal{I}m(\delta^d_{LL})_{23} &\sim \lambda \,, \quad \mathcal{I}m(\delta^d_{RR})_{23} \sim \lambda \,, \\ \mathcal{I}m(\delta^d_{LR})_{23} &\sim \lambda^4 \,, \quad \mathcal{I}m(\delta^d_{RL})_{23} \sim \lambda^4 \,. \end{aligned} \tag{183}$$

Examining Table 5, we learn that in alignment models $\mathcal{I}m(\delta^d_{LR})_{23} \sim 7 \times 10^{-4}(350\ \text{GeV}/\tilde{m})$ is the closest to satisfying the condition in Eq. (183), though the unknown numbers of order one should be on the large side to give an observable effect.

13.5 Discussion

We define two scales that play an important role in supersymmetry: Λ_S, where the soft supersymmetry breaking terms are generated, and Λ_F, where flavour dynamics takes place. When $\Lambda_F \gg \Lambda_S$, it is possible that there are no genuinely new sources of flavour and CP violation. This class of models, where the Yukawa couplings (or, in the mass basis, the CKM matrix) are the only source of flavour and CP breaking, is often called 'minimal flavour violation.' The most important features of the supersymmetry breaking terms are universality of the scalar masses-squared and proportionality of the A-terms. When $\Lambda_F \lesssim \Lambda_S$, we do not expect, in general, that flavour and CP violation are limited to the Yukawa matrices. One way to suppress CP violation would be to assume that, similarly to the SM, CP violating-phases are large, but their effects are screened, possibly by the same physics that explains the various flavour puzzles, such as models with Abelian or non-Abelian horizontal symmetries. It is also possible that CP-violating effects are suppressed because squarks are heavy. Another option, which is now excluded, was to assume that CP is an approximate symmetry of the full theory (namely, CP-violating phases are all small).

We would like to emphasize the following points:

1. For supersymmetry to be established, a direct observation of supersymmetric particles is necessary. Once it is discovered, then measurements of CP-violating observables will be a very sensitive probe of its flavour structure and, consequently, of the mechanism of dynamical supersymmetry breaking.
2. It seems possible to distinguish between models of exact universality and models with genuine supersymmetric flavour and CP violation. The former tend to give $d_N \lesssim 10^{-31}$ e cm, while the latter usually predict $d_N \gtrsim 10^{-28}$ e cm.
3. The proximity of $S_{\psi K_S}$ to the SM predictions is obviously consistent with models of exact universality. It disfavours models of heavy squarks such as that of Ref. [118]. Models of flavour symmetries allow deviations of order 20% (or smaller) from the SM predictions. To be convincingly signalled, an improvement in the theoretical calculations that lead to the SM predictions for $S_{\psi K_S}$ will be required [119].
4. Alternatively, the fact that $K \to \pi\nu\bar{\nu}$ decays are not affected by most supersymmetric flavour models [120–122] is an advantage here. The SM correlation between $a_{\pi\nu\bar{\nu}}$ and $S_{\psi K_S}$ is a much cleaner test than a comparison of $S_{\psi K_S}$ to the CKM constraints.
5. The neutral D system provides a stringent test of alignment. Observation of CP violation in the $D \to K\pi$ decays will make a convincing case for new physics.
6. CP violation in $b \to s$ transitions remains an interesting probe of supersymmetry. Deviations of order 0.1 from the SM predictions are possible if one of the conditions in Eq. (183) is satisfied.

14 Lessons from the B factories

Let us summarize the main lessons that have been learned from the measurements of CP violation in B decays:

- The KM phase is different from zero, that is, the SM violates CP.
- The KM mechanism is the dominant source of CP violation in meson decays.
- The size and the phase of new physics contributions to $b \to d$ transitions (B^0–$\overline{B}^0$ mixing) is severely constrained ($\leq \mathcal{O}(0.2)$).
- Complete alternatives to the KM mechanism (the superweak mechanism and approximate CP) are excluded.
- Corrections to the KM mechanism are possible, particularly for $b \to s$ transitions, but there is no evidence at present for such corrections.
- There is still a lot to be learned from future measurements.

Acknowledgements

I am grateful to Andreas Höcker, Sandrine Laplace and, in particular, Stephane T'Jampens for providing me with beautiful plots of CKM constraints. Their work has helped me to understand and, I hope, to explain the significance of the B-factory measurements of CP-violating asymmetries to our understanding of flavour and CP violation. I am grateful to Guy Raz and to Zoltan Ligeti for their contributions to the basic ideas and to the details of this review. I thank David Kirkby for collaboration on the PDG review on CP violation in meson decays [55], which is the basis of some sections in these lecture notes. This work was supported by a grant from the G.I.F., the German–Israeli Foundation for Scientific Research and Development, by the Israel Science Foundation founded by the Israel Academy of Sciences and Humanities, by EEC RTN contract HPRN-CT-00292-2002, by the Minerva Foundation (Munich), and by the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel.

References

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

[2] J. H. Christenson, J. W. Cronin, V. L. Fitch and R. Turlay, Phys. Rev. Lett. **13**, 138 (1964).

[3] H. Burkhardt *et al.* [NA31 Collaboration], Phys. Lett. B **206**, 169 (1988).

[4] V. Fanti *et al.* [NA48 Collaboration], Phys. Lett. B **465**, 335 (1999) [arXiv:hep-ex/9909022].

[5] A. Alavi-Harati *et al.* [KTeV Collaboration], Phys. Rev. Lett. **83**, 22 (1999) [arXiv:hep-ex/9905060].

[6] B. Aubert *et al.* [BaBar Collaboration], Phys. Rev. Lett. **87**, 091801 (2001) [hep-ex/0107013].

[7] K. Abe *et al.* [Belle Collaboration], Phys. Rev. Lett. **87**, 091802 (2001) [hep-ex/0107061].

[8] K. Abe [the Belle Collaboration], arXiv:hep-ex/0409049.

[9] B. Aubert [BaBar Collaboration], Phys. Rev. D **71**, 091102 (2005) [arXiv:hep-ex/0502019].

[10] T. Aushev [Belle Collaboration], arXiv:hep-ex/0411021.

[11] B. Aubert [BaBar Collaboration], Phys. Rev. Lett. **95**, 151804 (2005) [arXiv:hep-ex/0506082].

[12] B. Aubert [the BaBar Collaboration], Phys. Rev. Lett. **94**, 191802 (2005) [arXiv:hep-ex/0502017].

[13] K. Abe [Belle Collaboration], arXiv:hep-ex/0507037.

[14] B. Aubert [BaBar Collaboration], arXiv:hep-ex/0507087.

[15] B. Aubert *et al.* [BaBar Collaboration], arXiv:hep-ex/0408095.

[16] B. Aubert [BaBar Collaboration], Phys. Rev. Lett. **93**, 131801 (2004) [arXiv:hep-ex/0407057].

[17] K. Abe [Belle Collaboration], arXiv:hep-ex/0507045.

[18] C. C. Wang *et al.* [Belle Collaboration], Phys. Rev. Lett. **94**, 121801 (2005) [arXiv:hep-ex/0408003].

[19] B. Aubert *et al.* [BaBar Collaboration], arXiv:hep-ex/0408099.

[20] L. Wolfenstein, Phys. Rev. Lett. **13**, 562 (1964).

[21] G. Eyal and Y. Nir, Nucl. Phys. B **528**, 21 (1998) [arXiv:hep-ph/9801411].

[22] A. D. Sakharov, Pisma Zh. Eksp. Teor. Fiz. **5**, 32 (1967) [JETP Lett. **5**, 24 (1967)].

[23] G. R. Farrar and M. E. Shaposhnikov, Phys. Rev. D **50**, 774 (1994) [hep-ph/9305275].

[24] P. Huet and E. Sather, Phys. Rev. D **51**, 379 (1995) [hep-ph/9404302].

[25] M. B. Gavela, M. Lozano, J. Orloff and O. Pene, Nucl. Phys. B **430**, 345 (1994) [hep-ph/9406288].

[26] M. Fukugita and T. Yanagida, Phys. Lett. B **174**, 45 (1986).

[27] G. C. Branco, T. Morozumi, B. M. Nobre and M. N. Rebelo, hep-ph/0107164.

[28] A. G. Cohen, D. B. Kaplan and A. E. Nelson, Annu. Rev. Nucl. Part. Sci. **43**, 27 (1993) [hep-ph/9302210].

[29] R. J. Crewther, P. Di Vecchia, G. Veneziano and E. Witten, Phys. Lett. B **88**, 123 (1979) [Erratum-ibid. B **91**, 487 (1979)].

[30] L. J. Dixon, A. Langnau, Y. Nir and B. Warr, Phys. Lett. B **253**, 459 (1991).

[31] P. G. Harris *et al.*, Phys. Rev. Lett. **82**, 904 (1999).

[32] M. Dine, hep-ph/0011376.

[33] T. Banks, Y. Nir and N. Seiberg, hep-ph/9403203.

[34] R. D. Peccei and H. R. Quinn, Phys. Rev. Lett. **38**, 1440 (1977).

[35] R. D. Peccei and H. R. Quinn, Phys. Rev. D **16**, 1791 (1977).

[36] Y. Grossman and M. P. Worah, Phys. Lett. B **395**, 241 (1997) [hep-ph/9612269].

[37] Y. Grossman, G. Isidori and M. P. Worah, Phys. Rev. D **58**, 057504 (1998) [arXiv:hep-ph/9708305].

[38] G. Buchalla and A. J. Buras, Phys. Lett. B **333**, 221 (1994) [hep-ph/9405259].

[39] G. Buchalla and A. J. Buras, Phys. Rev. D **54**, 6782 (1996) [hep-ph/9607447].

[40] S. Bergmann and G. Perez, JHEP **0008**, 034 (2000) [hep-ph/0007170].

[41] M. P. Worah, Phys. Rev. Lett. **79**, 3810 (1997) [hep-ph/9704389].

[42] M. P. Worah, Phys. Rev. D **56**, 2010 (1997) [hep-ph/9702423].

[43] C. Jarlskog, Phys. Rev. Lett. **55**, 1039 (1985).

[44] M. C. Gonzalez-Garcia and Y. Nir, Rev. Mod. Phys. **75**, 345 (2003) [arXiv:hep-ph/0202058].

[45] N. Cabibbo, Phys. Rev. Lett. **10**, 531 (1963).

[46] Z. Maki, M. Nakagawa and S. Sakata, Prog. Theor. Phys. **28**, 870 (1962).

[47] L. Chau and W. Keung, Phys. Rev. Lett. **53**, 1802 (1984).

[48] L. Wolfenstein, Phys. Rev. Lett. **51**, 1945 (1983).

[49] A. J. Buras, M. E. Lautenbacher and G. Ostermaier, Phys. Rev. D **50**, 3433 (1994) [arXiv:hep-ph/9403384].

[50] C. Dib, I. Dunietz, F. J. Gilman and Y. Nir, Phys. Rev. D **41**, 1522 (1990).

[51] J. L. Rosner, A. I. Sanda and M. P. Schmidt, EFI-88-12-CHICAGO [presented at Workshop on High Sensitivity Beauty Physics, Batavia, IL, Nov 11–14, 1987].

[52] H. Harari and Y. Nir, Phys. Lett. B **195**, 586 (1987).

[53] S. Eidelman *et al.* [Particle Data Group], Phys. Lett. B **592**, 1 (2004).

[54] CKMfitter Group (J. Charles *et al.)*, Eur. Phys. J. **C 41**, 1–131 (2005), [hep-ph/0406184, updated results and plots available at: http://ckmfitter.in2p3.fr].

[55] D. Kirkby and Y. Nir, review of 'CP violation in meson decays' in Ref. [53].

[56] V. Weisskopf and E. P. Wigner, Z. Phys. **63**, 54 (1930); Z. Phys. **65**, 18 (1930) [see Appendix A of P. K. Kabir, "The CP Puzzle: Strange Decays of the Neutral Kaon", Academic Press (1968)].

[57] Y. Nir, SLAC-PUB-5874 [lectures given at 20th Annual SLAC Summer Institute on Particle Physics (Stanford, CA, 1992)].

[58] I. Dunietz and J. L. Rosner, Phys. Rev. D **34**, 1404 (1986).

[59] Ya. I. Azimov, N. G. Uraltsev and V. A. Khoze, Sov. J. Nucl. Phys. **45**, 878 (1987) [Yad. Fiz. **45**, 1412 (1987)].

[60] I. I. Bigi and A. I. Sanda, Nucl. Phys. B **281**, 41 (1987).

[61] Y. Grossman and Y. Nir, Phys. Lett. B **398**, 163 (1997) [arXiv:hep-ph/9701313].

[62] L. S. Littenberg, Phys. Rev. D **39**, 3322 (1989).

[63] A. J. Buras, Phys. Lett. B **333**, 476 (1994) [arXiv:hep-ph/9405368].

[64] G. Buchalla and A. J. Buras, Nucl. Phys. B **400**, 225 (1993).

[65] G. Blaylock, A. Seiden and Y. Nir, Phys. Lett. B **355**, 555 (1995) [hep-ph/9504306].

[66] A. F. Falk, Y. Grossman, Z. Ligeti and A. A. Petrov, Phys. Rev. D **65**, 054034 (2002) [arXiv:hep-ph/0110317].

[67] A. F. Falk, Y. Grossman, Z. Ligeti, Y. Nir and A. A. Petrov, Phys. Rev. D **69**, 114021 (2004) [arXiv:hep-ph/0402204].

[68] S. Bergmann and Y. Nir, JHEP **9909**, 031 (1999) [hep-ph/9909391].

[69] G. D'Ambrosio and D. Gao, Phys. Lett. B **513**, 123 (2001) [hep-ph/0105078].

[70] S. Bergmann, Y. Grossman, Z. Ligeti, Y. Nir and A. A. Petrov, Phys. Lett. B **486**, 418 (2000) [hep-ph/0005181].

[71] A. B. Carter and A. I. Sanda, Phys. Rev. Lett. **45**, 952 (1980).

[72] A. B. Carter and A. I. Sanda, Phys. Rev. D **23**, 1567 (1981).

[73] I. I. Bigi and A. I. Sanda, Nucl. Phys. B **193**, 85 (1981).

[74] G. Buchalla, A. J. Buras and M. E. Lautenbacher, Rev. Mod. Phys. **68**, 1125 (1996) [arXiv:hep-ph/9512380].

[75] Y. Grossman, A. L. Kagan and Z. Ligeti, Phys. Lett. B **538**, 327 (2002) [arXiv:hep-ph/0204212].

[76] H. Boos, T. Mannel and J. Reuter, Phys. Rev. D **70**, 036006 (2004) [arXiv:hep-ph/0403085].

[77] K. Abe, talk given at Lepton–Photon 2005.

[78] Y. Nir, Nucl. Phys. Proc. Suppl. **117**, 111 (2003) [arXiv:hep-ph/0208080].

[79] M. Beneke, G. Buchalla, M. Neubert and C. T. Sachrajda, Nucl. Phys. B **591**, 313 (2000) [hep-ph/0006124].

[80] G. Buchalla, G. Hiller, Y. Nir and G. Raz, JHEP **0509**, 074 (2005) [arXiv:hep-ph/0503151].

[81] A. Ali, G. Kramer and C. D. Lü, Phys. Rev. D **58**, 094009 (1998) [arXiv:hep-ph/9804363].

[82] M. Beneke and M. Neubert, Nucl. Phys. B **651**, 225 (2003) [arXiv:hep-ph/0210085].

[83] M. Beneke, Phys. Lett. B **620**, 143 (2005) [arXiv:hep-ph/0505075].

[84] H. J. Lipkin, Phys. Rev. Lett. **46**, 1307 (1981).

[85] M. Gronau and D. London, Phys. Rev. Lett. **65**, 3381 (1990).

[86] Y. Grossman and H. R. Quinn, Phys. Rev. D **58**, 017504 (1998) [hep-ph/9712306].

[87] J. Charles, Phys. Rev. D **59**, 054007 (1999) [hep-ph/9806468].

[88] M. Gronau, D. London, N. Sinha and R. Sinha, Phys. Lett. B **514**, 315 (2001) [hep-ph/0105308].

[89] H. J. Lipkin, Y. Nir, H. R. Quinn and A. Snyder, Phys. Rev. D **44**, 1454 (1991).

[90] M. Gronau, Phys. Lett. B **265**, 389 (1991).

[91] A. E. Snyder and H. R. Quinn, Phys. Rev. D **48**, 2139 (1993).

[92] H. R. Quinn and J. P. Silva, Phys. Rev. D **62**, 054002 (2000) [arXiv:hep-ph/0001290].

[93] A. F. Falk, Z. Ligeti, Y. Nir and H. Quinn, Phys. Rev. D **69**, 011502 (2004) [arXiv:hep-ph/0310242].

[94] M. Gronau and D. London, Phys. Lett. B **253**, 483 (1991).

[95] M. Gronau and D. Wyler, Phys. Lett. B **265**, 172 (1991).

[96] D. Atwood, I. Dunietz and A. Soni, Phys. Rev. Lett. **78**, 3257 (1997) [arXiv:hep-ph/9612433].

[97] A. Giri, Y. Grossman, A. Soffer and J. Zupan, Phys. Rev. D **68**, 054018 (2003) [arXiv:hep-ph/0303187].

[98] Y. Grossman, Y. Nir and M. P. Worah, Phys. Lett. B **407**, 307 (1997) [hep-ph/9704287].

[99] Y. Nir and D. J. Silverman, Nucl. Phys. B **345**, 301 (1990).

[100] K. Agashe, M. Papucci, G. Perez and D. Pirjol, arXiv:hep-ph/0509117.

[101] Y. Grossman, Y. Nir and R. Rattazzi, hep-ph/9701231.

[102] M. Dine, E. Kramer, Y. Nir and Y. Shadmi, Phys. Rev. D **63**, 116005 (2001) [hep-ph/0101092].

[103] H. E. Haber, Nucl. Phys. Proc. Suppl. **62**, 469 (1998) [hep-ph/9709450].

[104] M. Dugan, B. Grinstein and L. Hall, Nucl. Phys. B **255**, 413 (1985).

[105] S. Dimopoulos and S. Thomas, Nucl. Phys. B **465**, 23 (1996) [hep-ph/9510220].

[106] W. Buchmuller and D. Wyler, Phys. Lett. B **121**, 321 (1983).

[107] J. Polchinski and M. B. Wise, Phys. Lett. B **125**, 393 (1983).

[108] W. Fischler, S. Paban and S. Thomas, Phys. Lett. B **289**, 373 (1992) [hep-ph/9205233].

[109] F. Gabbiani, E. Gabrielli, A. Masiero and L. Silvestrini, Nucl. Phys. B **477**, 321 (1996) [hep-ph/9604387].

[110] Y. Nir, Nucl. Phys. B **273**, 567 (1986).

[111] G. Eyal, A. Masiero, Y. Nir and L. Silvestrini, JHEP **9911**, 032 (1999) [arXiv:hep-ph/9908382].

[112] D. Becirevic *et al.*, Nucl. Phys. B **634**, 105 (2002) [arXiv:hep-ph/0112303].

[113] M. Ciuchini, E. Franco, G. Martinelli, A. Masiero, M. Pierini and L. Silvestrini, arXiv:hep-ph/0407073.

[114] L. Silvestrini, arXiv:hep-ph/0510077.

[115] Y. Nir and N. Seiberg, Phys. Lett. B **309**, 337 (1993) [arXiv:hep-ph/9304307].

[116] M. Leurer, Y. Nir and N. Seiberg, Nucl. Phys. B **420**, 468 (1994) [arXiv:hep-ph/9310320].

[117] Y. Nir and G. Raz, Phys. Rev. D **66**, 035007 (2002) [arXiv:hep-ph/0206064].

[118] A. G. Cohen, D. B. Kaplan, F. Lepeintre and A. E. Nelson, Phys. Rev. Lett. **78**, 2300 (1997) [hep-ph/9610252].

[119] G. Eyal, Y. Nir and G. Perez, JHEP **0008**, 028 (2000) [hep-ph/0008009].

[120] Y. Nir and M. P. Worah, Phys. Lett. B **423**, 319 (1998) [hep-ph/9711215].

[121] A. J. Buras, A. Romanino and L. Silvestrini, Nucl. Phys. B **520**, 3 (1998) [hep-ph/9712398].

[122] G. Colangelo and G. Isidori, JHEP **9809**, 009 (1998) [hep-ph/9808487].

Experimental aspects of cosmic rays

P. Sommers
University of Utah, Salt Lake City, UT 84112-0830, USA

Abstract
High-energy cosmic rays are detected as extensive air showers, and properties of the primary cosmic rays are deduced from measurements of those air showers. The physics of air showers is reviewed here in order to explain how the measurement techniques work. The Pierre Auger Cosmic Ray Observatory (near this school in Malargue) is used to illustrate the experimental methods. The Auger Observatory combines a surface array of water Cherenkov detectors with atmospheric fluorescence detectors. This ‘hybrid’ measurement technique provides high resolution and measurement cross-checks. In conjunction with a complementary site in the northern hemisphere, the Auger Observatory expects to map the arrival directions over the full sky as well as measuring the cosmic-ray energy spectrum and statistical properties of the mass distribution.

1 Introduction

A high-energy cosmic ray observatory records individual cosmic ray particles using the atmosphere as a transducer and amplifier. Each extremely high energy cosmic ray converts into a cascade that grows to billions of secondary particles. A large observatory records the air showers landing in a collection area that spans thousands of square kilometres. Because of the indirect measurement method, it is impossible to measure exactly the arrival direction, energy, and mass of the primary cosmic ray. Air shower measurements can determine the arrival direction to a small fraction of a degree. The energy, however, is hard to measure to better than 10% accuracy. The atomic mass of the primary particle cannot be estimated reliably for individual cosmic rays, and only some statistical properties of the primary mass distribution can be derived from air shower studies.

The flux of high-energy cosmic rays is tiny. Above 10^{19} eV, for example, the detection rate is approximately one per square kilometre per year. The rate falls by two orders of magnitude for each decade increase in particle energy. Direct measurement of the highest energy cosmic rays above the atmosphere is far from feasible. The study of extremely high energy cosmic rays must rely on indirect measurements via the air showers that those particles produce. Modern (hybrid) observatories combine a surface array of particle detectors with telescopes that observe radiation produced by the developing shower front as it traverses the atmosphere.

Physicists have used air showers to study cosmic rays near 10^{14} eV and above since Pierre Auger demonstrated the technique in 1938 [1]. The largest air-shower array, located near Malargue, is the Pierre Auger Observatory. It is designed for the study of the highest energy cosmic rays, making quality measurements of all air showers above 10^{19} eV that land within its 3000 km^2 area. Its surface detector (SD) is an array of 1600 water Cherenkov tanks. On good-weather nights, its air fluorescence detector (FD) measures the longitudinal development of the cascade as it descends through the atmosphere. The Auger Observatory will be used to illustrate the experimental techniques that are employed to measure the highest energy cosmic rays.

2 Air shower physics

An extremely high energy cosmic ray collides with a nucleus high in the atmosphere. The interaction produces many new energetic particles. Those also collide with air nuclei, and each collision adds a large number of particles to the developing cascade. Some of the produced particles are neutral pions, each one of which immediately decays to a pair of gamma rays. The gamma rays produce $e^{\pm}$ pairs when passing near nuclei. The electrons and positrons re-generate gamma rays via bremsstrahlung, thereby building the electromagnetic cascade. This is an extensive air shower.

The number of charged particles in the air shower reaches a maximum size N_{max} that is nearly proportional to the primary energy E. There are billions of charged particles in high-energy air showers. The size N_{max} at shower maximum is approximately equal to E/1.5 GeV, although this conversion factor depends slightly on the choice of hadronic interaction model that is adopted to simulate collisions at energies above the reach of collider experiments, and it depends slightly on the atomic mass of the cosmic ray.

Experimental evidence so far indicates that extremely high energy cosmic rays are atomic nuclei (including protons which are hydrogen nuclei). The cosmic-ray nucleus initiates its air-shower cascade by hadronic interaction with an atomic nucleus in the atmosphere. The hadronic cascade (mostly pions) grows until the energy per pion falls to the level where pions are likely to decay before colliding. In each generation of the hadronic cascade, 1/3 of the energy on average goes to neutral pions which instantly decay to pairs of gamma rays. Each gamma ray develops an electromagnetic subcascade. After n hadronic cascade generations, only $(2/3)^n$ of the total energy remains in the hadronic cascade.

The decay of π^0 mesons into gamma rays eventually transfers most of the primary cosmic ray's energy to the electromagnetic cascade. Each gamma ray converts to an $e^{\pm}$ pair. The electrons and positrons create new gamma rays by bremsstrahlung. The radiation length X_{r} is the grammage path length in which their energies attenuate by the factor $1/e$. In air, this radiation length X_{r} is 36.2 g/cm^2. The electromagnetic cascade grows via pair production and bremsstrahlung.

Heitler's heuristic picture [2] of the electromagnetic cascade gives intuitive understanding of its essential properties. One imagines the cascade developing by a sequence of generations. At each generation, any gamma ray produces an $e^{\pm}$ pair, while each electron or positron produces a gamma ray in addition to itself. Every generation therefore doubles the number of cascade particles. The grammage interval for each generation is $X_{\text{r}}\times$ ln(2), i.e., the path over which the energy of any one particle is expected to be reduced by 1/2. The process continues until the average particle energy is reduced to the *critical energy* below which charged particles lose their energy in less than one radiation length by ionizing atoms. Given that the ionization energy loss is about 2.2 MeV/g/cm^2, the critical energy E_{c} is (2.2 MeV/g/cm^2)×(36.2 g/cm^2) = 80 MeV. The cascade grows until it reaches size N$_{\text{max}}$ $= E/E_{\text{c}}$. The number of generations n needed to reach this maximum size depends on the total energy E. Since the number of particles doubles at each generation, one has at maximum, $N_{\text{max}} = 2^n = E/E_{\text{c}}$, so $n =$ ln(E/E_{c})/ln(2). The maximum size occurs at a slant depth $X_{\text{max}} = n \times X_{\text{r}} \times$ ln(2) $=X_{\text{r}} \times$ ln(E/E_{c}) (measured along the shower axis from the top of the atmosphere).

Rigorous treatments show that this heuristic model gives the correct depth of maximum (X_{max}) for each energy. In particular, the depth of maximum X_{max} for electromagnetic cascades increases by $X_{\text{r}}\times$ ln(10) for each decade of increase in energy E. This *elongation rate* of 85 g/cm^2/decade is greater than what is expected for air showers that are fed by hadronic cascades.

The heuristic model's suggestion that N_{max} is proportional to energy is only approximately true. The total energy is equal to 2.2 MeV times the charged particle shower size integrated over all depths, as the electromagnetic energy is dissipated by charged particles at the rate of 2.2 MeV/g/cm^2/particle. Because the *longitudinal profile* $N_{\text{e}}(X)$ gets longer (X_{max} larger) with energy, the height of the profile curve cannot be strictly proportional to energy. In fact, $N_{\text{max}} \propto E/\sqrt{\text{ln}(E)}$. To a good approximation, this is proportional to E in accordance with the heuristic model.

Rigorous treatments show that the electromagnetic longitudinal profile is accurately given by the Greisen formula [3]:

$$N_{\text{e}} = \frac{0.31}{\sqrt{T_{\text{max}}}}\, e^T\, s^{-3T/2}\ .$$

Here T is the atmospheric depth measured in radiation lengths ($T = X/X_\mathrm{r}$) from the point of the gamma ray's production, $T_\mathrm{max} \equiv \ln(E/E_\mathrm{c})$, and s is the *shower age*: $s \equiv 3T/(T + 2T_\mathrm{max})$. Many shower properties are well parametrized by shower age. All positive values of atmospheric depth occur in the range $0 < s < 3$, and shower maximum occurs at shower age $s = 1$.

In the hadronic cascade, the charged pions produce air shower muons when they decay. The number of shower muons depends on the amount of energy that is left in the hadronic cascade when pion energies have dropped to the level where decay is more likely than collision. If this happens after relatively few cascade generations, then copious muon production occurs. If the reduction of pion energies takes relatively many generations, then more of the energy will have been lost from the hadronic cascade to the electromagnetic cascade, and meager muon production occurs.

A cascade initiated by an iron nucleus develops like a superposition of 56 nucleons, each with 1/56 of the primary energy. In effect, this jump-starts the cascade, and pions get down to energies where they can decay to muons before the electromagnetic cascade has drained too much energy from the hadronic cascade. An iron shower therefore typically has more muons than a proton shower of the same total energy. Moreover, the superposition of 56 lower energy subshowers reaches its maximum size higher in the atmosphere than a proton shower of the same total energy. Statistical determinations of the primary mass distribution (chemical composition) exploit these differences between heavy and light nuclei: heavy nucleus showers produce more muons and they reach maximum size higher in the atmosphere.

The longitudinal development (rise and fall of the number of charged particles) is frequently approximated by a parametrized Gaisser–Hillas functional form [4]:

$$N_\mathrm{e}(x) = N_\mathrm{max}\, (\frac{x}{w})^w\, e^{w-x} \,,$$

where $w \equiv (X_\mathrm{max} - X_0)/\lambda$, $x \equiv (X - X_0)/\lambda$, and λ is an interaction scale length in g/cm^2. The four parameters (N_max, X_max, X_0, λ) provide ample size and shape freedom for fitting longitudinal profiles. As an exercise, try fitting the Greisen function (given above) by a Gaisser–Hillas function with appropriate choice of parameters.

3 Geometric reconstruction

Surface arrays determine the arrival direction by recording the arrival time of the shower front at three or more non-collinear stations on the ground. The method is conceptually simple assuming the shower front to be a perfect plane. Any pair of stations A and B determine the arrival direction cosine along the direction from A to B as $c(t_\mathrm{A} - t_\mathrm{B})/\bar{\mathrm{AB}}$, where c is the speed of light, t_A and t_B are the trigger times for stations A and B, respectively, and $\bar{\mathrm{AB}}$ is the distance between them. Two independent direction cosines determine a unique arrival direction in the hemisphere above the plane of the detectors.

The shower front is actually a curved surface, not a plane. The trigger times of three or more stations give the geometry by chi-square minimization, using the expected relative arrival times based on a realistic curved shower front moving at the speed of light. Those expected times depend not only on the arrival direction but also on the core position, so the core should first be determined from the relative station signal amplitudes.

Geometric reconstruction is quite different for a fluorescence detector eye, which sees an air shower as a spot of light that moves downward through the atmosphere at the speed of light. The track of the spot's centre defines a great circle in the direction space of the eye which, together with the eye's location, determines the shower-detector plane (SDP). If two eyes at different locations record the same shower, then the shower axis must lie in both SDPs. The intersection of the planes determines the shower axis (provided the two planes are not the same). This is the *stereo* reconstruction method. The angular resolution depends on the accuracy of determining the SDPs and on the opening angle between them. The SDP accuracy is better for longer tracks and smaller pixels [5].

Data from a single FD eye together with the trigger time(s) from one or more SD stations yield a hybrid geometric reconstruction that offers better accuracy than stereo reconstruction. After determining the SDP, 'geometric reconstruction' means identifying the shower axis within the SDP together with the time when the shower front passes some point on that axis. To understand the hybrid method, imagine that you know precisely the angular velocity of the track as the spot passes the centre of some particular pixel. If somebody were to tell you the distance to the shower axis at that point of the track, you could calculate when the light was emitted from that point of the axis and (using the measured angular velocity) what angle the axis makes with the pixel's viewing direction. The geometric reconstruction is therefore complete if you are told that one distance to the axis. Since you may not know it accurately from the FD data alone, you can try all possible distance hypotheses. Each one gives a unique geometric reconstruction and therefore a unique prediction for the trigger time of any SD station on the ground. The trigger time of any ground station therefore identifies which distance hypothesis is correct and therefore the true geometry. This timing method typically picks out the axis within the SDP with less uncertainty than if the axis is determined by a second (stereoscopic) SDP. For stereo hybrid events, there are timing determinations in two independent planes, providing reconstruction accuracy that is superior to both stereo reconstruction and monocular hybrid reconstruction.

4 Energy measurement

Conceptually, energy determination by a fluorescence detector is straightforward. The amount of emitted fluorescence light is proportional to the ionization energy loss by all the charged particles. Measuring the fluorescence emission from the full shower development should yield the total electromagnetic shower energy. It is a robust calorimetric measurement. The only dependence on hadronic model or composition is in the small fraction of primary energy that is assumed to escape the hadronic cascade as muons and neutrinos rather than being transferred to the electromagnetic cascade. Simulations suggest that this fraction is approximately 5% for proton showers and 15% for iron showers. By assuming 10%, the error due to ignorance of the primary particle should not be more than about 5%. Still, it is important to recognize that air shower measurements are an indirect method for determining the energy of a cosmic ray, and there is *some* systematic model uncertainty that is difficult to quantify. The fraction of the cosmic ray energy not dissipated electromagnetically also fluctuates shower-to-shower, especially for protons.

Implementing this conceptually simple calorimetric method encounters numerous difficulties:

1. The full longitudinal development is never observed. The FD records only the portion of the shower development that is above ground level and large enough to produce a detectable light flux at the detector. Some extrapolation using a fitted functional form is needed to account for the parts of the development that are not measured.
2. Light scattered to the FD from the intense forward Cherenkov beam contaminates the fluorescence signal. This scattered Cherenkov light distorts the spectrum of detected photons as well as their number.
3. The optical clarity of the atmosphere is variable because of changes in the aerosol density and aerosol composition. This makes it problematic to infer the amount of emitted light based on the observed flux. Detailed atmospheric monitoring can, in principle, overcome this difficulty.
4. The fluorescence efficiency is not precisely known. The number of fluorescence photons produced per metre along a charged particle's trajectory depends on the particle's energy and also on the atmospheric temperature and pressure where the particle is. Uncertainty in the fluorescence efficiency causes uncertainty in inferring energy deposition based on the amount of produced fluorescence light.

Numerous laboratory experiments are tackling the last itemized difficulty [6], and they can be expected to reduce the uncertainty in cosmic-ray energy measurements which is due to uncertainty in the fluorescence yield. The other difficulties introduce energy errors that are more random than systematic.

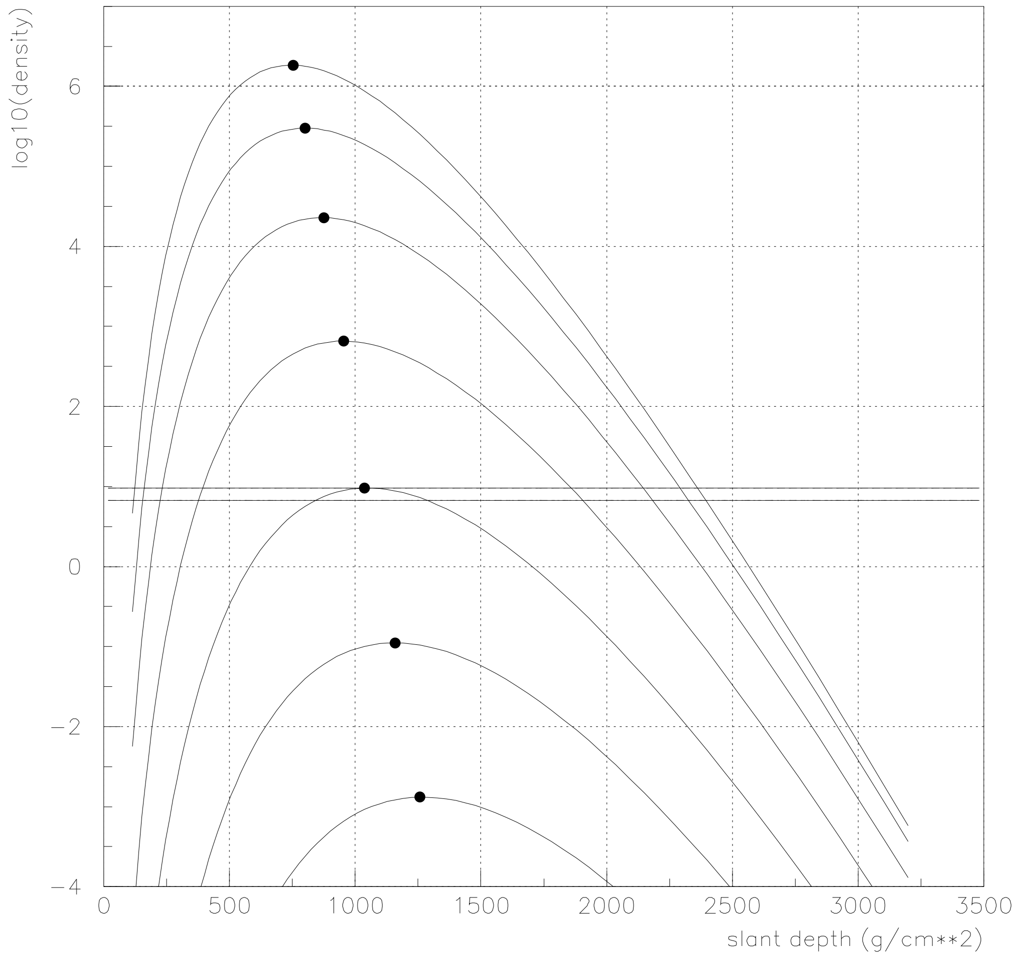


Fig. 1: Longitudinal development curves for the NKG electromagnetic particle density at seven fixed distances from the shower axis. The distances are fixed in Molière units but converted to metres using the Auger detector altitude. Top to bottom, the distances (in surface metres) to the axis are 10, 32, 100, 320, 1000, 3200, 10 000. Note that the depth of maximum increases with distance from the axis. Two horizontal lines mark the maximum of s_{1000} curve and the value 30% lower. The density remains within 30% of its maximum from 850 to 1300 g/cm^2. The units are approximate number of vertical equivalent muons per 10 m^2 for this shower of $N_{\rm max} = 6 \times 10^9$.

The FD quasi-calorimetric energy measurement can provide an important calibration for SD energy measurements which otherwise rely on shower simulations. Shower simulations are necessarily uncertain in their treatment of hadronic interactions at energies that have not been studied by collider experiments.

Simulations show that the signal collected in SD stations far from the core is approximately proportional to the total shower energy. In the case of the Auger array, the signal (s_{1000}) deposited in a water tank 1000 meters from the core is taken to be proportional to shower energy. At that distance, the longitudinal development of particle density reaches its maximum value near ground level for a large range of zenith angles. Since a smooth function changes very little near its maximum value, this method is relatively insensitive to fluctuations in shower development.

Figure 1 shows the longitudinal development profiles for the particle density at seven different distances from the shower axis. The curves are analytic: the total shower size as a function of depth is taken to be a Gaisser–Hillas development curve and the lateral distribution at each depth (hence shower age) is given by the NKG function [7, 8]:

$$\rho(r) = \frac{N(s)}{R_{\rm M}^2}\, r^{s-2}\,(1+r)^{s-4.5} \;/\; [2\pi B(s, 4.5-2s)] \,,$$

where $N(s)$ is the number of charged particles at age s and $B(x, y) \equiv \Gamma(x)\Gamma(y)/\Gamma(x+y)$ is the standard 'beta function'. The different curves correspond to different fixed distances from the axis measured in Molière units, using the single conversion to metres that pertains at the Auger ground level. Each density longitudinal profile has a dot indicating its maximum. You can see that the depth of maximum increases with Molière distance from the core. Although the total shower size reaches maximum at 800 g/cm^2 in this example, the density s_{1000} (measured at the Molière radius that corresponds to one kilometre at ground level) reaches maximum at more than 1000 g/cm^2. A detector at vertical depth 850 g/cm^2 (Auger) or 920 g/cm^2 (AGASA) is therefore well positioned to be near the s_{1000} maximum for zenith angles at least up to 45 degrees. The horizontal dashed lines in the figure mark the maximum of the s_{1000} curve and the value that is 30% less. You can also see that the density stays within 30% of its maximum from 850 g/cm^2 to 1300 g/cm^2. Fluctuations in shower development (and even the systematic difference between protons and iron) shift the maximum by only about 100 g/cm^2 or less.

Knowing the slant depth of a shower's measurement, the expected longitudinal development profile of s_{1000} allows one to correct the measured s_{1000} to what it likely was at maximum. For large zenith angles, these correction factors become dangerously large for scintillator arrays. For water Cherenkov detectors, the longitudinal profiles fall much slower with slant depth than the (electromagnetic) NKG behaviour plotted in Fig. 1. For water Cherenkov arrays, therefore, relatively small corrections to s_{1000} are needed even at large zenith angles to give the ground parameter (s_{1000} at its maximum), which is proportional to energy.

None of the triggered ground stations is likely to be exactly 1000 metres from the axis. Interpolation using numerous stations closer to and farther from the axis gives the estimate for the signal density at 1000 m. Rather than using linear interpolation, one fits an empirical average lateral distribution functional form and takes as s_{1000} the value of that fitted function at 1000 metres from the axis.

A statistical correlation of s_{1000} with shower energy measured calorimetrically by the FD yields the conversion factor from s_{1000} to energy without relying on shower simulations that use untested hadronic interaction models. A large hybrid data set will determine the conversion factor's dependence on zenith angle.

5 Composition analysis

The indirect measurement of a cosmic ray by its air shower makes it impossible to measure the mass of the primary particle exactly. Shower-to-shower fluctuations in longitudinal development can make an air shower produced by a particle of one mass indistinguishable from an air shower initiated by a particle of a different mass. As described in Section 2, however, each mass value leads to its own expected value for the shower depth of maximum and number of muons (at each energy). These quantities are both intimately related to the speed of shower development, but fluctuations in $X_{\rm max}$ are not highly correlated with fluctuations in N_μ. For example, a fluctuation in depth of first interaction relates directly to a shower's $X_{\rm max}$, but it has little impact on the shower's N_μ. Direct or indirect measurements of $X_{\rm max}$ and/or N_μ provide important information about the primary mass distribution, even though it is not feasible to determine the masses of individual primary cosmic rays.

The 'superposition model' is a useful approximation that is remarkably accurate (although not exactly correct). Its premise is that *an air shower by a nucleus (E,A) behaves like the superposition of A proton showers, each with energy E/A.* (Note that this is not the same as supposing that it breaks into A nucleons at its first interaction.) The expected difference between proton and iron air showers of a given energy can be evaluated simply using this superposition model. Their difference in $X_{\rm max}$ and number of muons N_μ will be considered here.

Denoting the expected value of X_{max} by $\bar{X}_{\text{max}}$, the difference between proton and iron at fixed energy E is given by

$$\bar{X}^{\text{P}}_{\text{max}}(E) - \bar{X}^{\text{Fe}}_{\text{max}}(E) = \bar{X}^{\text{P}}_{\text{max}}(E) - \bar{X}^{\text{P}}_{\text{max}}(E/56) = ER \times \text{Log}(56) \; .$$

Here ER is the proton elongation rate (change in X_{max} per decade change in energy). Taking $ER \sim 57$ g/cm^2/decade implies that proton air showers are expected to reach maximum development approximately 100 g/cm^2 deeper in the atmosphere than iron showers of the same total energy.

To evaluate the difference in muon content N_μ one assumes a power-law dependence of $\bar{N}_\mu$ on energy for proton showers: $\bar{N}_\mu(E) = \alpha E^\beta$. For this argument, this power-law dependence only needs to be a good approximation over an energy range $E/56$ to E for the energy of interest. With this assumption, then,

$$\bar{N}^{\text{Fe}}_\mu(E) = 56\bar{N}^{\text{P}}_\mu(E/56) = 56\alpha(E/56)^\beta = 56^{1-\beta}\alpha E^\beta = 56^{1-\beta}\bar{N}^{\text{P}}_\mu(E) \; .$$

Therefore,

$$\bar{N}^{\text{Fe}}_\mu(E)/\bar{N}^{\text{P}}_\mu(E) = 56^{1-\beta} \; .$$

At energies near 10^{19} eV, $\beta \approx 0.93$, so this ratio is approximately 1.3. Iron showers are expected to have 30% more muons than proton showers of the same total energy.

A surface detector measures shower parameters that are sensitive to X_{max}. These include the lateral steepness and the shower front curvature. A shower whose maximum is farther above the surface has a flatter lateral distribution and less curvature in its shower front. The shape of the detector signal pulses can be sensitive to the relative contribution from muons. Muons suffer less Coulomb scattering and arrive more promptly than electromagnetic particles. A shower by a heavy nucleus (like iron) is expected to produce signal pulses that rise and fall rapidly relative to pulses in showers produced by light nuclei (like protons). Cronin [9] has therefore advocated studying composition by the distribution of the 'shape parameter' of signal pulses in water Cherenkov detectors at a distance of 1000 metres from the core. The suggested parameter is the ratio of signal accumulated in the first 600 ns to signal collected after the first 600 ns. Heavy primaries are indicated by large values of the shape parameter and light primaries by small values.

Fluorescence detectors can measure X_{max} directly. Analyses of the mean X_{max} as a function of energy in Fly's Eye [10] and HiRes [11, 12] data have suggested that the composition gets significantly lighter with energy near or before the spectrum's ankle.

A hybrid data set has special value in studying the cosmic ray chemical composition. The FD measures the electromagnetic shower energy and the depth of maximum X_{max}. Water Cherenkov detectors are sensitive to the muon composition. At any fixed electromagnetic energy, heavy and light components can be separated when s_{1000} is plotted versus X_{max}, as shown in Fig. 2 [13].

Combining also a scintillator array with a water Cherenkov array and fluorescence detectors should allow an even more powerful probe of the primary mass distribution. The scintillator array is sensitive to the surface electromagnetic particle density, allowing that to be subtracted from the water Cherenkov signal to measure the contribution due to muons.

Determining the primary mass distribution is challenging because of the indirect method of measuring cosmic rays by their air showers. There are many measurable parameters that correlate with the primary mass, however. These include the depth of shower maximum X_{max}, the shower front curvature, the signal rise times, and the steepness of the lateral distribution. Multi-parameter analyses may provide the best sensitivity. Neural networks are a special kind of multiparameter analysis. A neural net can provide a mass likelihood distribution for each measured shower, based on its multiparameter training with simulated showers.

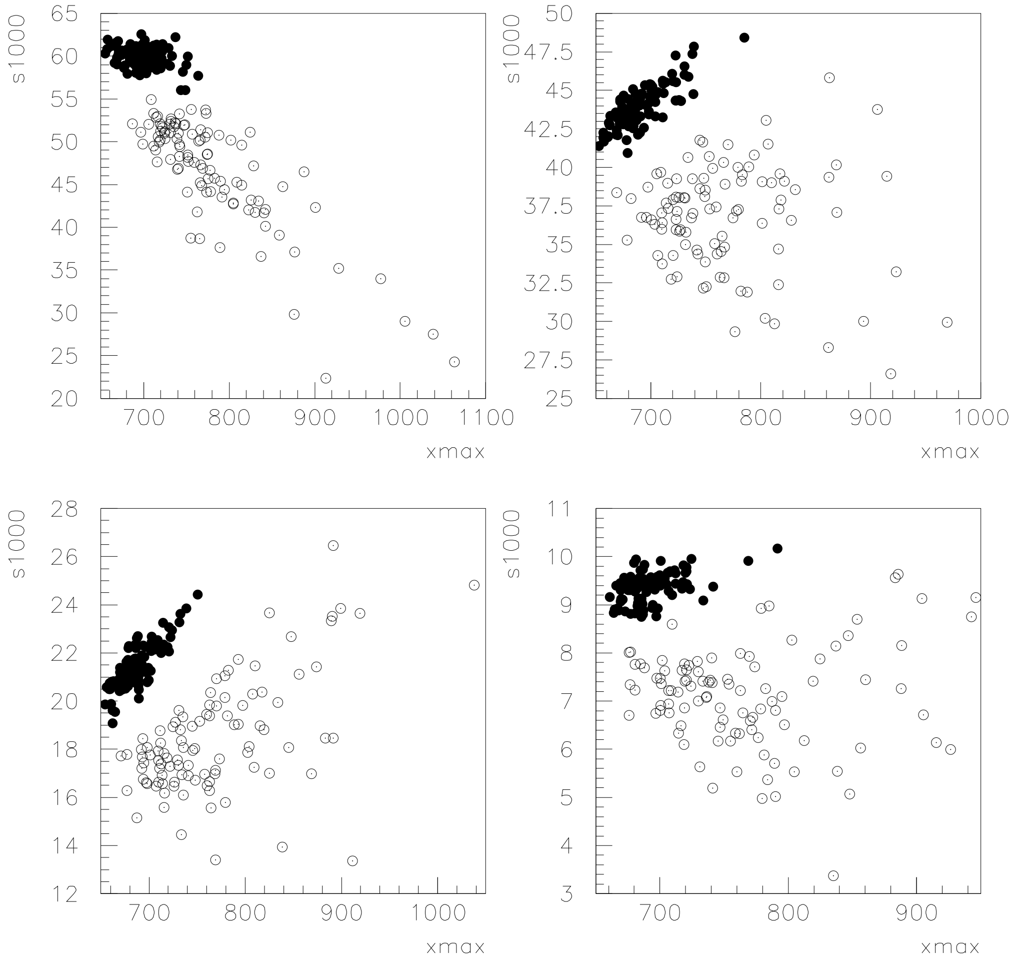


Fig. 2: Proton showers (○) are separated from iron showers (●) in this scatter plot of s_{1000} vs. X_{max}. These are simulated showers of 10^{19} eV at zenith angles of 0, 37, 53, and 66 degrees, using CORSIKA and QGSJET.

6 Identifying photon primaries

Extremely high energy gamma-ray primary particles produce electromagnetic air showers that differ significantly from those produced by primary nuclei. Some muons occur by virtue of pion photoproduction, but the number of muons is an order of magnitude less than in showers initiated by primary hadrons. Moreover, the depth of maximum is expected to be greater for a gamma-ray primary than for a primary hadron. Recall that the 'elongation rate' for electromagnetic showers is 85 g/cm^2/decade, whereas it is only about 57 g/cm^2/decade for protons. (High-multiplicity hadronic interactions serve to divide the primary energy faster than the electromagnetic pair-production and bremsstrahlung processes.) Although proton showers develop similarly to gamma-ray showers at much lower energies, the proton showers develop significantly faster near 10^{19} eV and above.

The techniques for studying chemical composition using direct or indirect measures of X_{max} or muon content are able to detect the existence or not of a component of gamma rays in the extremely high energy cosmic rays. The Cronin shape parameter, shower front curvature, lateral steepness, and direct measurement of X_{max} should all be able to identify gamma-ray primaries if they are present in the cosmic ray flux. Some models for cosmic ray production (especially the 'top-down' models in which the cosmic rays result from the decays of supermassive particles) predict a dominant (or at least

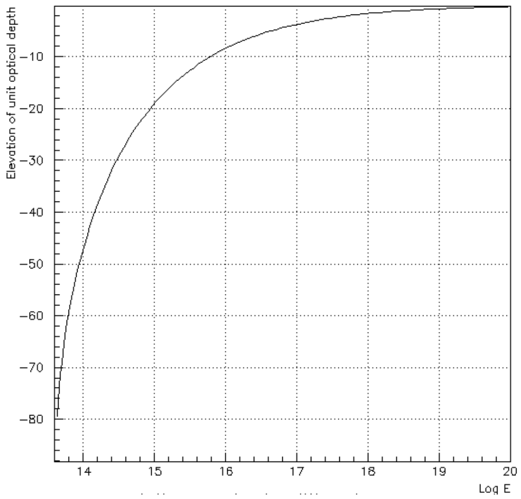


Fig. 3: The elevation angle (negative) which corresponds to optical depth of 1 for neutrinos coming through the Earth as a function of neutrino energy. The Earth is nearly opaque above 10^{17} eV.

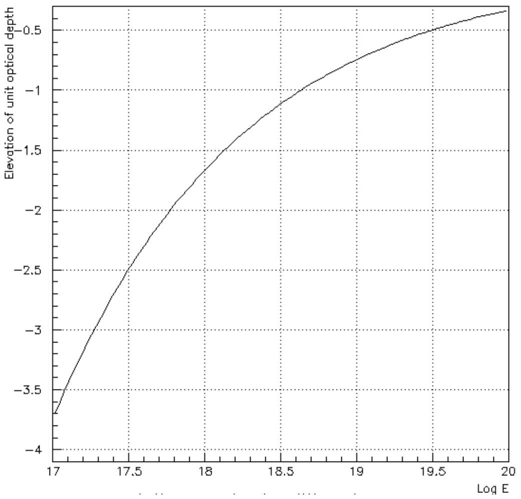


Fig. 4: Detail of the left figure for energy greater than 10^{17} eV. Above 10^{18} eV, the neutrinos must only skim the Earth at an upward angle less than 2 degrees from horizontal.

significant) component of primary gamma rays at extremely high energy. The absence of a clear gamma-ray component would rule out many such scenarios.

Above 10^{19} eV, two other effects introduce a different signature of a gamma-ray component. One effect is the Landau–Pomeranchuk–Migdal (LPM) effect [14], which decreases the electromagnetic cross sections and consequently lengthens the electromagnetic shower development. The other effect is magnetic pair production in the Earth's magnetosphere [15]. A superhigh-energy gamma ray crossing magnetic field lines can produce an $e^{\pm}$ pair. Those will synchrotron-radiate in the magnetic field, producing additional high-energy gamma rays. A different kind of electromagnetic cascade develops in the magnetosphere, but the effect is similar to adding some radiation lengths above the atmosphere. The important thing is that this occurs only if the primary gamma ray arrives transverse to the geomagnetic field. If it arrives approximately parallel to field lines, then there is no 'pre-showering'. Gamma rays arriving transverse to field lines cascade to lower energy particles above the atmosphere and are little affected by the LPM effect. Their atmospheric longitudinal developments are shortened by their head-start in the magnetosphere. In contrast, those same gamma rays arriving parallel to the magnetic field lines would get the full LPM effect and have stretched-out shower developments. Looking at shower development speed as a function of arrival direction relative to the local magnetic field direction is a way to look for evidence (or not) of gamma rays at the highest energies.

7 Neutrino detection

A giant cosmic-ray observatory like Auger makes an effective neutrino observatory at EeV energies. The detection is especially efficient for tau neutrinos, so it is fortunate that neutrino mixing makes tau neutrinos approximately as abundant as the other flavours for astrophysical high-energy neutrinos. An EeV tauon lives long enough to travel 50 km, on average. A tauon produced in the ground with a trajectory of small elevation angle will decay into showering particles above the surface array, and the air shower will be sampled by the particle detectors. The signature of a young (small age) electromagnetic shower is clearly distinguishable from a (very old) near-horizontal shower produced by a normal cosmic ray. The shower front curvature is much smaller, and the individual detector flash-analog-to-digital converter (FADC) traces are much broader.

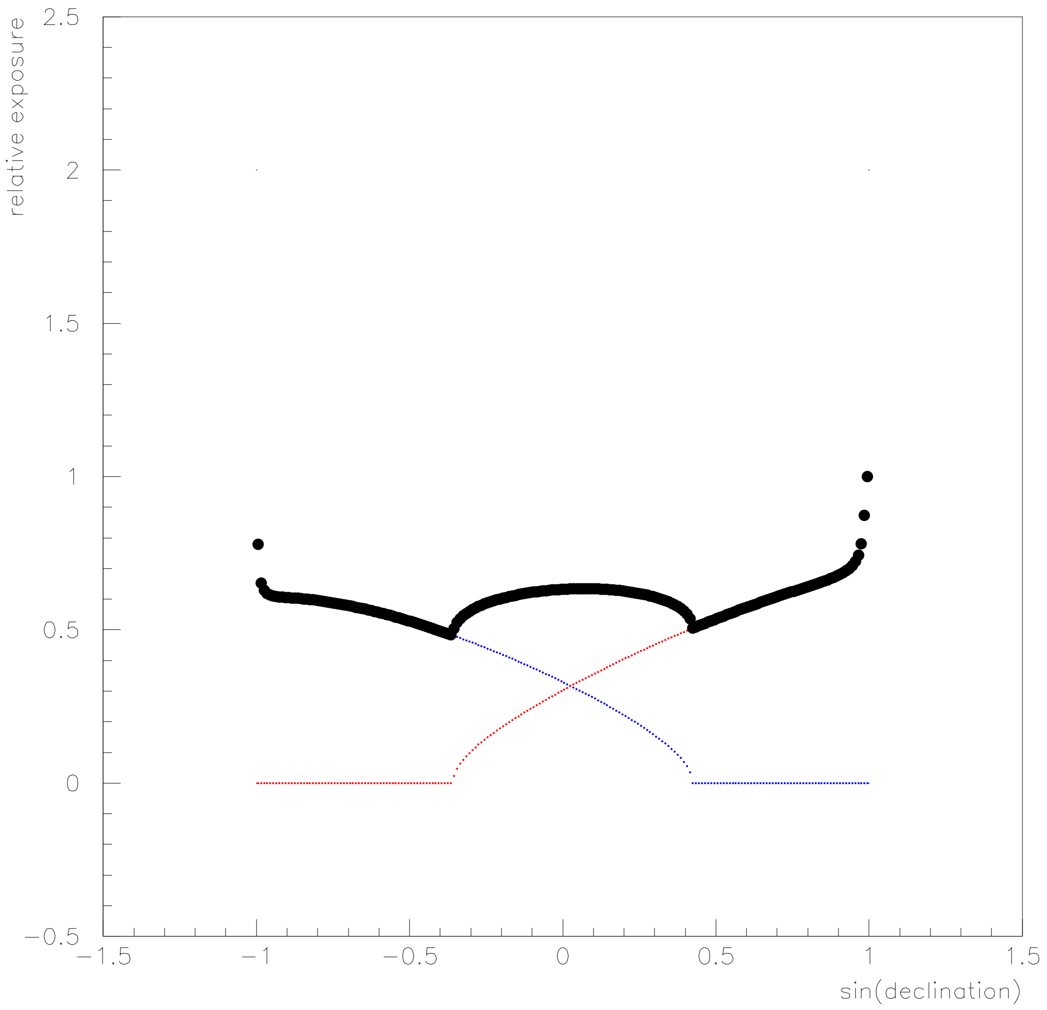


Fig. 5: The exposure versus celestial declination. This is for an observatory array with acceptance out to 60 degrees in zenith angle. The southern site is assumed to be at latitude –35 degrees and the northern site at +39 degrees.

The Earth is opaque to neutrinos at EeV energies and above, as shown in Figs. 3 and 4. This means that detectable neutrinos must come from near-horizontal 'Earth-skimming' directions, so that the total grammage of their trajectory in the Earth is not too large. This does not reduce the Auger aperture (as it does for detectors like IceCube) because Auger can only detect and distinguish air showers resulting from neutrino interactions if they are nearly horizontal. Auger should be able to detect some neutrinos arising from the GZK pion photoproduction process [16]. The magnitude of this neutrino flux at EeV energies is an important handle on the production and propagation of the highest energy cosmic rays.

8 Anisotropy analysis

The search for the sources of high-energy cosmic rays requires measuring more than their energy spectrum and composition. Those do not adequately constrain the possible theories. Whatever spectrum and composition might be measured, theorists would find multiple models for the origin of cosmic rays that are compatible with those results. An anisotropy fingerprint is needed to identify the sources definitively. This is true whether the sources are resolved as discrete spots on the sky or whether their signature is a large-scale pattern.

The Auger Observatory was designed to detect cosmic rays with almost uniform sensitivity over the entire celestial sphere. Because the Earth is opaque to cosmic rays (and rotates only about one axis), full-sky exposure requires detectors in both the northern and southern hemispheres. Figure 5 shows that the southern site by itself is blind to 1/4 of the sky, and there is a steep gradient over half of the remainder. Similarly, the northern site is blind to the 1/4 of the sky that the southern site sees well, and its gradient is opposite that of the southern site in the 1/2 of the sky where they overlap. Their combination produces nearly uniform sensitivity to the entire sky.

It is interesting to note that the blind spot of the southern site includes the richest concentration of matter in the nearby universe (out to 21 Mpc) [17]. It is therefore dangerous to suppose that the cosmic rays observed at the southern site are a fair representation of the full sky. The energy spectrum, the composition, and the clustering of arrival directions could all be different in the north and south. It is essential to make high-statistics measurements with identical detector types in both hemispheres.

Multipole moments are the natural way to characterize celestial anisotropy, provided full-sky coverage is available [18]. Each moment is the coefficient a_{lm} of a spherical harmonic function $Y_{lm}(\theta, \phi)$ in a series expansion of the celestial function. The lowest order harmonics (small l-values) govern the large-scale structure (dipole, quadrupole, octupole moments). The moments with large l-values determine the fine structure. Expansions out to $l \sim 30$ are ample for cosmic-ray measurements with angular resolution on the order of a degree. Magnetic dispersion may make it pointless to go beyond about $l = 10$. (At the highest energies where the magnetic dispersion is small, there will be too few arrival directions to determine many multipole moments.)

An example of an anisotropy fingerprint up to $l = 10$ is shown in Fig. 6. This heuristic example is based on an artificial assumption that the cosmic-ray sources are nearby infrared sources given in the point source catalogue PSCz [19]. The number of arrival directions from each source is taken in proportion to its apparent infrared luminosity. The magnetic dispersion is done using the source distance and a simple model of uniform nanogauss magnetic fields with coherence length of 1 Mpc. A total of 36 000 simulated arrival directions were used. They produced the multipole moments that are the fingerprint in Fig. 6. A simulation of an isotropic flux is also plotted in the figure for comparison. The anisotropy is manifest, and the fingerprint characterizes it. The fingerprint defines a function on the celestial sphere. Sampling arrival directions with that weighting function produces a scatter plot that is fully consistent with the original.

A rich celestial pattern like this is not possible for a realistic cosmic-ray observatory. Above the GZK energy threshold, it is possible to expect charged-particle astronomy. Particles are not deflected much by magnetic fields, and the sources cannot be far away. The nearby sources should be identifiable by the cosmic-ray arrival directions. Unfortunately, however, practical observatories cannot hope to achieve the exposure needed to get thousands of super-GZK arrival directions.

The 36 000 directions of Fig. 6 would be appropriate for a 5-year combined exposure of Auger South and Auger North at energies above 10^{19} eV, well below the GZK threshold. In that case, the majority of arrival directions will be cosmic rays that were produced billions of years ago, and which have been fully isotropized. The anisotropic foreground sources account for only about 1/60th of all arrival directions in that case. It is therefore relevant to ask if the fingerprint of Fig. 6 could be detected when it is diluted 59-to-1 by an isotropic distribution. As shown in Fig. 7, the answer is Yes. Any distribution of 36 000 arrival directions can be tested against the fingerprint of Fig. 6. The method is to compute its multipole moments (a'_{lm}) and then sum the products:

$$S \equiv \Sigma_{lm} \left(a_{lm} a'_{lm}\right) .$$

(This is like integrating the product of the two celestial functions. If they are similar, then the integral picks up systematically positive values rather than summing positive and negative values equally.) As shown in Fig. 7, simulations with 1/60th of the arrival directions being sampled from the distribution of Fig. 6 (and 59/60 isotropically) give a product sum that is distinctly non-zero. Simulations of complete

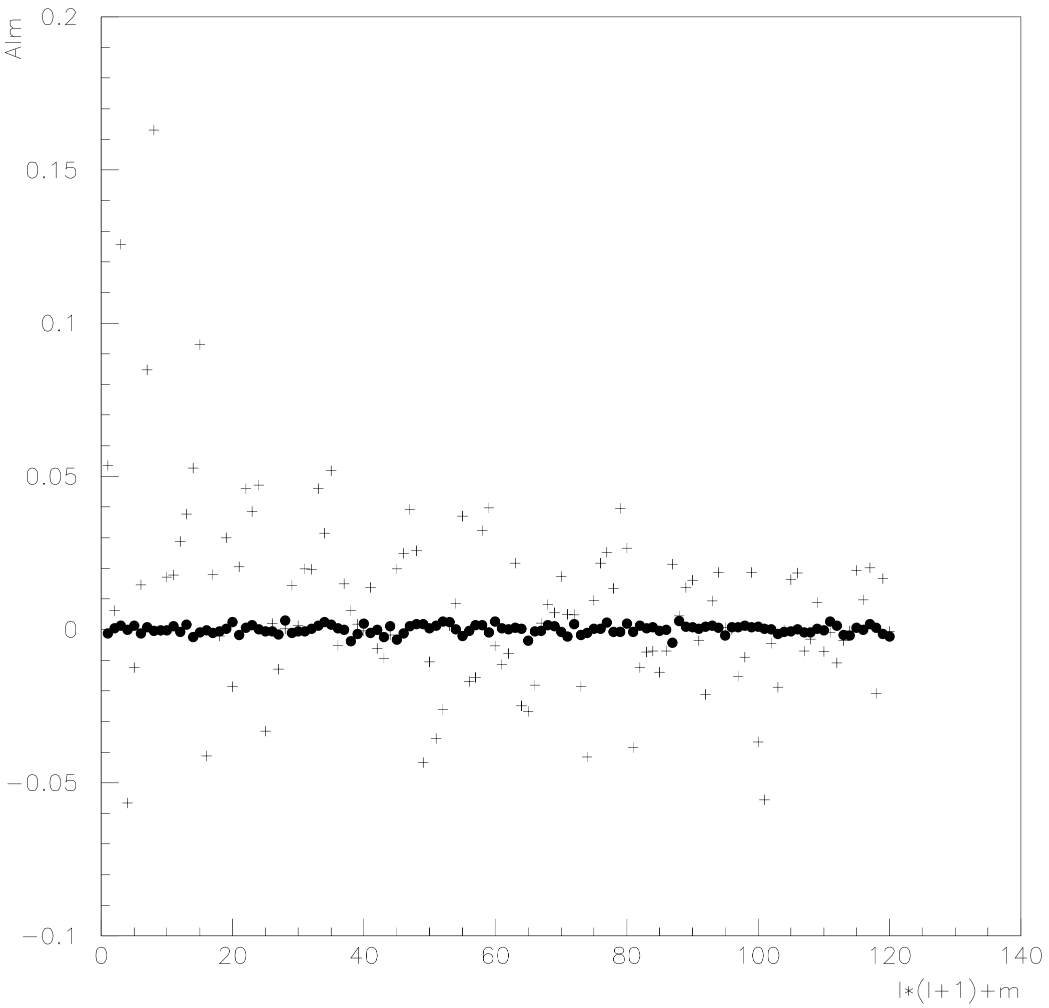


Fig. 6: An example of an anisotropy fingerprint. The 121 a_{lm} coefficients for $l \leq 10$ are plotted for a specific simple model of cosmic rays coming from nearby infrared sources. These are plotted with + signs. Also shown with • marks are results for an isotropic simulation with the same number of arrival directions (36 000). The set of a_{lm} values in the anisotropic case constitute a fingerprint of the anisotropy.

isotropy do not have a systematic offset from zero. The two distributions are almost disjoint, indicating that the foreground sources in this artificial model can be identified by their known fingerprint even in the presence of the dominant isotropic background.

Multipoles are not defined for a function on just part of a sphere. Without full-sky coverage, a cosmic-ray observatory cannot determine any of the a_{lm} coefficients. The powerful fingerprinting via multipole moments requires a full-sky observatory.

9 Summary

Properties of high-energy cosmic rays are inferred from measurements of their air showers. The cosmic-ray arrival direction, energy, and mass are determined indirectly. Combining different air-shower measurement techniques is important since each technique by itself is susceptible to systematic error as well as shower-by-shower uncertainties that result from shower development fluctuations and also fluctuations in the detector samplings.

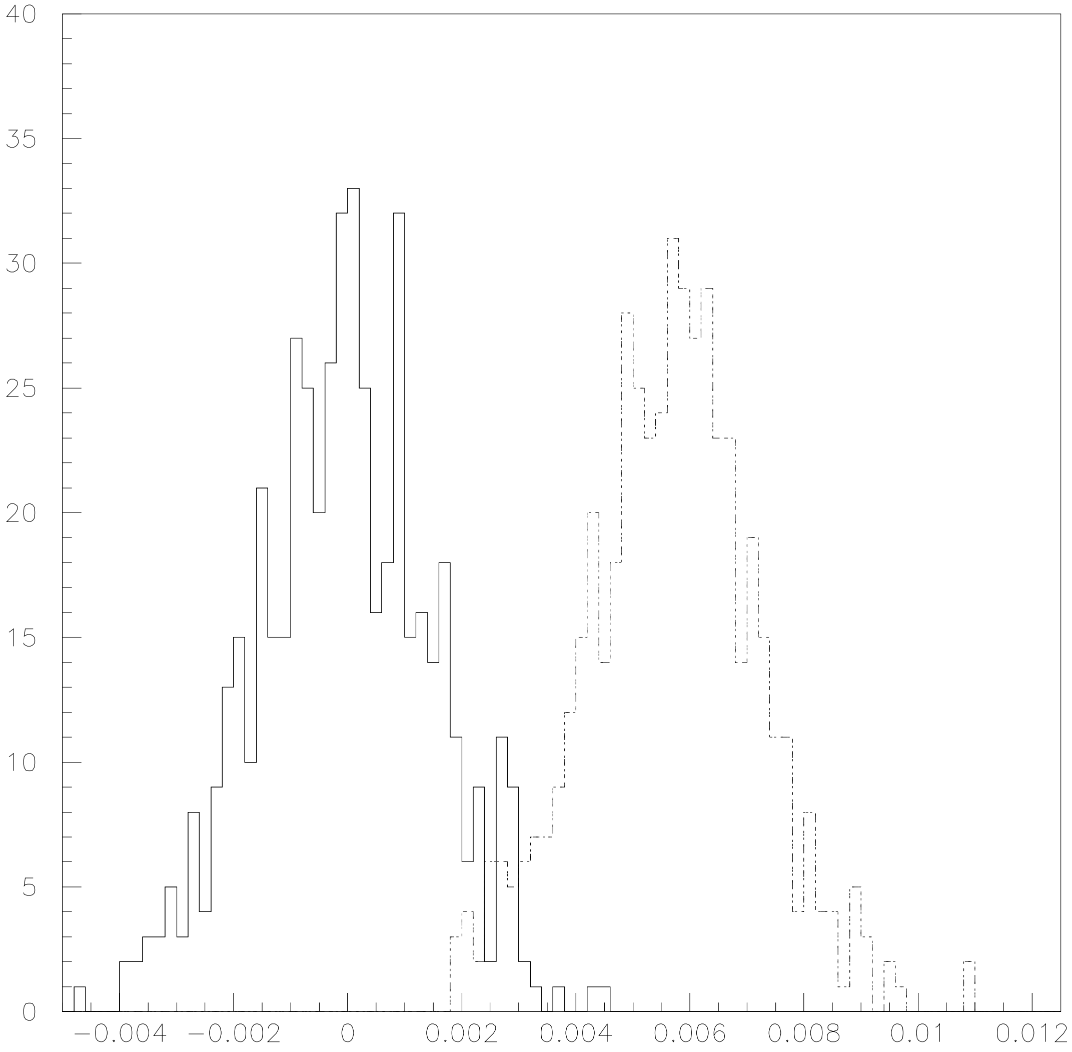


Fig. 7: Distinguishing the diluted anisotropy pattern. The left histogram (centred on zero) is derived from isotropic simulations, whereas, for the right histogram, each simulation has 1/60th of the arrival directions sampled from the PSCz anisotropy pattern used in Fig. 6. In each case, the variable represented in the histogram is the sum (over l and m) of the products of a_{lm} from the simulation times the a_{lm} of the PSCz anisotropy represented in Fig. 6.

The Auger Observatory, for example, combines a surface array of water Cherenkov detectors with telescopes that record air fluorescence produced by the secondary shower particles as they descend through the atmosphere. The hybrid air-shower measurements are crucial. They provide precision angular reconstructions for detector resolution studies and point source searches. They provide precision core locations for the study of SD-only geometric reconstructions. They determine the relationship of s_{1000} to electromagnetic energy at different zenith angles and energies. They provide X_{max} and the electromagnetic energy together with shower front curvature, lateral distribution steepness, and SD rise times for use in composition studies.

The surface detector operates on its own 90% of the time. It provides the high statistics needed for the high-energy end of the spectrum. It will measure the minuscule super-GZK flux. It has good sensitivity to EeV neutrinos. Its exposure is nearly uniform in right ascension due to the Earth's rotation. The declination dependence is easily calculated from the zenith angle acceptance. So the celestial exposure has a simple form which is easily calculated. Together with the high SD statistics, this results in sensitive anisotropy measurements.

Anisotropy analyses using data from only the southern site will be handicapped by a steep exposure gradient over half of the sky and a total blind spot in the quarter of the sky that is richest in nearby matter concentrations. The combination of Auger South with Auger North is needed for truly sensitive full-sky anisotropy analyses. A multipole moment 'fingerprint' could then characterize the celestial anisotropy in a small table of numbers.

References

[1] P. Auger *et al., Comptes Rendus* **206** (1938) 1721.

[2] M.A. Longair, *High Energy Astrophysics* (Cambridge University Press, 1992), vol. 1, pp. 118–122; W. Heitler, *The Quantum Theory of Radiation* (Dover, NY, 1984).

[3] K. Greisen, *Prog. Cosmic Rays* **3** (1956) 1.

[4] T. Gaisser and M. Hillas, *Proc. 15th ICRC*, Plovdiv, Bulgaria **8** (1977) 353.

[5] J.W. Elbert, Track reconstruction techniques and geometrical resolution in HiRes, *Proc. Tokyo Workshop on Techniques for the Study of Extremely High Energy Cosmic Rays*, M. Nagano [Ed.], (Institute for Cosmic Ray Research, University of Tokyo, 1993), p.158.

[6] http://www.auger.de/events/air-light-03/.

[7] K. Greisen, *Annu. Rev. Nucl. Sci* **10** (1960) 63.

[8] K. Kamata and J. Nishimura, *Prog. Theor. Phys.* (Kyoto) Suppl. 6 (1958) 93.

[9] J.W. Cronin, Particle discrimination using the FADC traces from the Auger Observatory surface detectors, Auger technical note GAP-2003-076 (2003).

[10] T.K. Gaisser *et al.,* A study of the chemical composition of cosmic rays around 10^{18} eV, *Comments Astrophys.* **17** (1993) 103–115.

[11] T. Abu-Zayyad *et al.,* Measurement of the cosmic ray spectrum and composition form 10^{17} to 10^{18} eV using a hybrid fluorescence technique, *Astrophys. J.* **557** (2001) 686.

[12] T. Abu-Zayyad *et al.,* A multi-component measurement of the cosmic ray composition between 10^{17} eV and 10^{18} eV, astro-ph/9911144.

[13] P. Billoir, *Comptes Rendus* **5** (2004) 495.

[14] L.D. Landau and I.Ya. Pomeranchuk, *Dokl. Akad. Nauk USSR* **92** (1953) 535; A.B. Migdal, *Phys. Rev.* **103** (1956) 1811.

[15] B. McBreen and C.J. Lambert, *Proc. 17th Int. Cosmic Ray Conf.* **6** (Paris, 1981) 70; T. Erber, *Rev. Mod. Phys.* **38** (1966) 626.

[16] K. Greisen, *Phys. Rev. Lett.* **16** (1966) 748; G.T. Zatsepin and V.A. Kuz'min, *JETP Lett.* **4** (1966) 78.

[17] J.W. Cronin, The highest-energy cosmic rays, Taup 2003 proceedings, astro-ph/0402487.

[18] P. Sommers, Cosmic ray anisotropy analysis with a full-sky observatory, *Astropart. Phys.*, **14** (2001) 271.

[19] http://www-astro.physics.ox.ac.uk/wjs/pscz.html

The Pierre Auger Observatory – why and how

A.A. Watson
School of Physics and Astronomy, University of Leeds, UK

1 Some historical background

Cosmic rays were discovered by an intrepid Austrian balloonist, Victor Hess, in 1912. In a remarkable series of balloon flights, one of which took him above 5000 m, he showed that the rate of formation of ions in a closed chamber increased with altitude. He concluded that the Earth was being bombarded by radiation from outer space which was given the name 'cosmic radiation (or cosmic rays)' by R.A. Millikan in 1926. He thought that γ rays, then the most penetrating radiation known, caused the enhanced ion production observed by Hess.

We now know that only about 10^{-4} of the incoming radiation is in the form of γ rays. Most of the radiations are atomic nuclei with about 1% primary electrons. The energy range extends from about 1 GeV (where the solar magnetic field can deflect the particles) to at least 10^{20} eV. Because the particles are charged, and interstellar and intergalactic space are threaded with magnetic fields, it has not been possible to trace the particles back to their point of production. Above about 10^{14} eV the flux of cosmic rays is so low that it is barely practical to detect them directly using instruments carried on balloons or spacecraft and instead one must rely on the extensive air showers (EAS) that the particles create when they hit the Earth's atmosphere. Above 10^{14} eV, where the maximum number of $\sim 10^5$ particles is reached at $\sim$ 6 km above sea-level, some particles survive so that remnants of the primary are detectable. Because of scattering, electrons and photons can be found at large distances from the axis of such showers although about 50% lie within the Molière radius which is about 70 m at sea-level. The discovery of extensive air showers is usually credited to Pierre Auger [1] who, in 1938, observed an unexpectedly high rate of coincidences between counters separated by a few metres. Further investigations by his team showed that even when the counters were as far as 300 m apart, the rate of coincidence was significantly in excess of the chance expectation. Speculating that the primaries were photons, and using the newly developed ideas of quantum electrodynamics, Auger demonstrated that the incoming entities had energies as high as 10^{15} eV. Earlier Rossi [2] had reported experimental evidence for extensive groups of particles ("sciami molto estesi di corpscoli") which produced coincidences between counters rather distant from each other. Kolhörster and colleagues [3] made very similar observations to those of Auger and his group with counter separations out to 75 m. It was Auger, however, who was in a position to follow up the discovery of this new phenomenon and through his inferences about the primaries extend the range of energies then known by nearly 6 orders of magnitude. Cosmic rays remain the most extreme example of the departure of matter from thermal equilibrium.

One of the early motivations for studying cosmic rays using extensive air showers was the expectation that anisotropies would be discovered as the technique allowed the exploration of an energy regime where deflections by a galactic magnetic field might be small enough to permit the observation of point sources. This led to the construction of larger and larger shower arrays where 'large' eventually meant an area of a few square kilometres. Such detectors were developed at Volcano Ranch, USA (8 km^2, with scintillators), Haverah Park, UK (12 km^2, with water-Cherenkov detectors), SUGAR, Australia ($\sim$ 100 km^2, with buried scintillators) and Yakutsk, Siberia (25 km^2, with scintillators, muon detectors, and air-Cherenkov detectors). In 1963 Linsley reported the detection of an event of 10^{20} eV (or 100 EeV) with the Volcano Ranch array [4]. The significance of this energy was not immediately appreciated but soon after the discovery of the 2.7 K cosmic microwave radiation in 1966, Greisen and Zatsepin and Kuz'min [5] pointed out that if the highest energy particles were protons and if their sources were uniformly distributed throughout the Universe, then there would be interactions between the cosmic rays and the microwave background that would modulate the spectrum of the highest energy particles.

A particularly important reaction is the following:

$$\mathrm{p} + \gamma_{2.7\mathrm{K}} \rightarrow \Delta^{+}(1232) \rightarrow \mathrm{p} + \pi^{0} \quad \text{or} \quad \mathrm{n} + \pi^{+} \, . \tag{1}$$

In the rest frame of the proton (the cosmic ray), the microwave background photon will look like a very-high-energy γ ray. When the Lorentz factor Γ of the proton is $\sim 10^{11}$, the Δ^{+} resonance will be excited. In each reaction (1) the proton loses about 15% of its energy. Over cosmological distances, sufficient reactions take place for the observed spectrum to become significantly depleted of ultra-high-energy cosmic rays (UHECRs) compared with what might have been present at the time of acceleration. It follows that if protons of $\sim 10^{20}$ eV are observed, they must have originated from nearby. For example, a cosmic ray of 50 EeV has a 50% chance of having come from beyond 100 Mpc. This opens the prospect of seeing point sources of cosmic rays at the very highest energies as the intergalactic magnetic fields are not expected to bend the trajectories of the particles by too large an amount.

If high-energy particles can escape from the acceleration region as nuclei, then the CMB radiation, supplemented by the diffuse infrared radiation field, are important factors. The key resonance is now the giant dipole resonance (typical energy $\sim$ 10 MeV), and the mixture of species that arrives at the Earth can be very complex depending upon the paths travelled through the radiation fields. Both protons and nuclei also lose energy by pair production, the threshold here corresponding to $\Gamma \sim 10^{9}$. The energy losses are small but nearly continuous and may be important if protons of energies 1–10 EeV are of extragalactic origin. The reactions of Eq. (1) are also important in the context of γ and ν astronomy while the neutrons from photodisintegration are also a source of neutrinos.

2 Detectors and measurements from the pre-Auger era

In Fig. 1 a representation of the cosmic-ray energy spectrum due to Gaisser [6] is displayed. Nearly all of the data shown above 10^{14} eV are from air-shower measurements. It is relevant in the context of the CERN–CLAF School to point out the position of the Tevatron and the LHC on the energy axis. The arrows indicate the energies that a cosmic ray hitting a stationary nucleon must have for the centre-of-mass energy to be the same as achieved in a Tevatron or LHC collision.

Thus it is evident that knowledge of the particle physics interactions, even from the LHC, will not cover the energy range of relevance to cosmic rays of the highest energy. Furthermore, the region of rapidity space that will be observed at the LHC (Fig. 2) excludes the diffractive region that is of great importance in the development of an air shower. In a shower, the energy carried by the leading particle from each collision, which may be $\sim$ 0.5 of the incoming energy, is crucial for the development of the shower, just as the multiplicity of the charged meson component radiating from the collision is crucial to the development of the muon signal. Neutral pions play a key role in the development of the electromagnetic cascade.

It follows that significant extrapolation is required to infer what has initiated an air shower from what is observed at ground-level. Ideally, knowledge of the mass and of the hadronic physics is required at the energies of interest, where the hadronic physics must cover details of pion–nucleus collisions and nucleus–nucleus collisions at extreme energies.

In the 1970s an alternative technique to that of deploying particle detectors over greater and greater areas emerged. This relies on the excitation of atmospheric nitrogen by the electrons of the shower as it traverses the atmosphere. The nitrogen emits fluorescence radiation isotropically, predominantly in the 300–400 nm band, and this can be observed at distances of $\sim$ 20 km with arrays of photomultipliers on dark, clear nights. The technique was pioneered by a group from the University of Utah. With their original Fly's Eye detector they recorded an event of 300 EeV, still the highest cosmic-ray energy ever claimed. This event is shown in Fig. 3(a) together with a schematic of a photomultiplier array [Fig. 3(b)] in which each tube is orientated in a different direction. The magnitudes of the signals in the tubes are used to estimate the number of particles along the track of the shower while the positions of the tubes

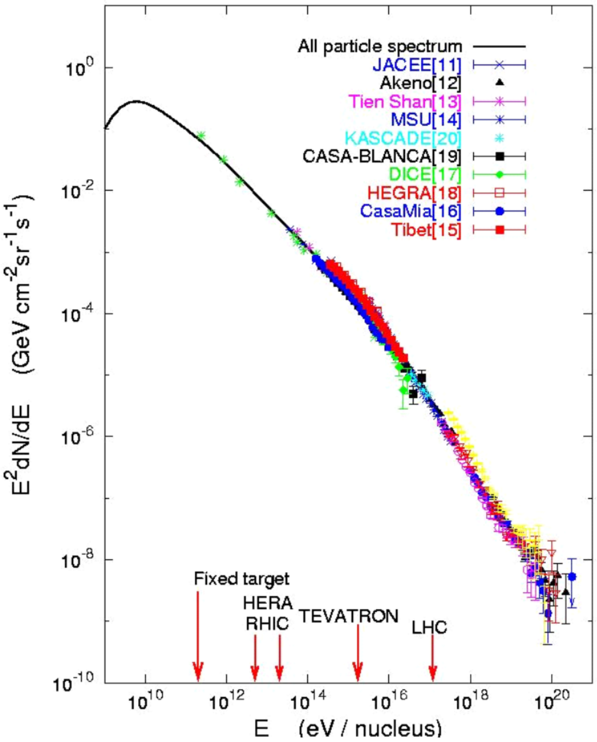


Fig. 1: Data summary made by Gaisser [6]. Below about 10^{14}eV it is possible to make observations directly in spacecraft or, after correcting for the atmospheric overburden, from balloons. Above this energy the data are deduced almost exclusively from studies of extensive air showers. The spectrum is rather featureless: the marked bend at around 1 GeV is caused by the solar magnetic field; there is a small steepening in the spectrum at about 3 × 10^{15} eV (known as the knee) and around 3 × 10^{18} eV the spectrum flattens again at the ‘ankle’. What the details are above 10^{19} eV (or 10 EeV) remains to be resolved.

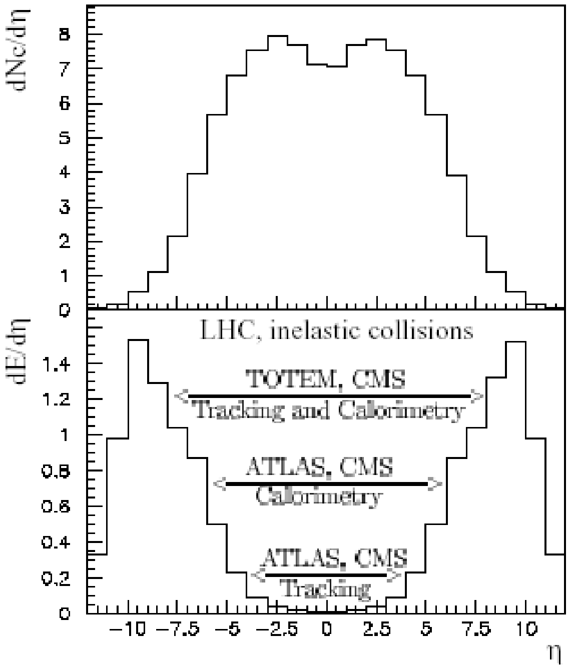
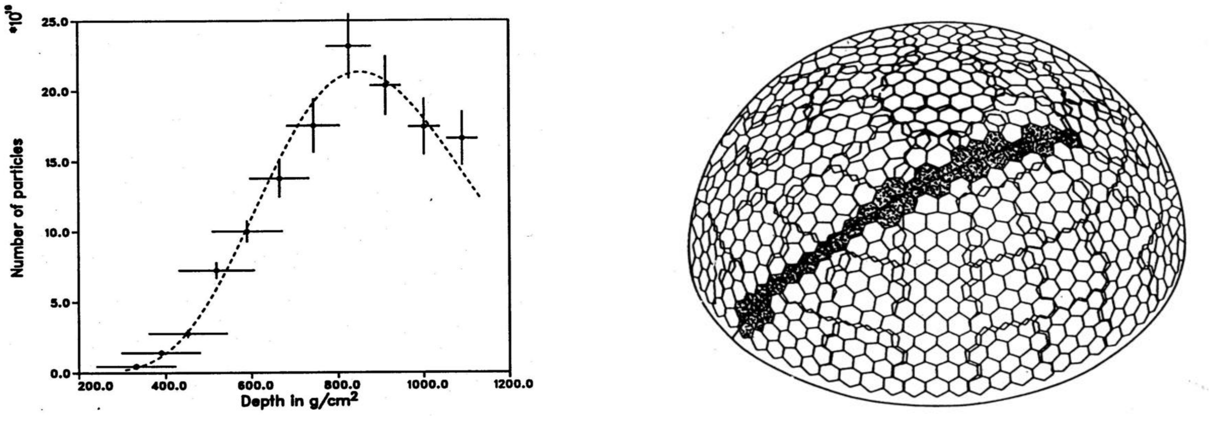


Fig. 2: The pseudo-rapidity (η) distribution of charged particles (upper panel) and of the energy flow (lower panel) predicted for pp collisions at the LHC [7]

define a plane in which the axis of the shower lies (the shower detector plane). The primary energy is then derived by integrating under the longitudinal development curve (the track length integral) and multiplying the result by the appropriate (-dE/dx). This gives a calorimetric estimate of the total energy deposited in the atmosphere.



(a)

(b)

Fig. 3: (a) The longitudinal development of a shower created by a primary cosmic ray of 300 EeV as recorded by the Fly's Eye detector [8]; (b) A schematic diagram illustrating a shower crossing the array of photomultiplier tubes in the Fly's Eye detector.

To find the energy of the primary particle, this estimate of the ionization energy loss must be augmented by about 7–10% to allow for energy carried by high-energy muons and neutrinos into the ground. This correction is slightly dependent on the mass and hadronic interaction model assumed but has a much smaller systematic uncertainty than has the conversion to primary energy from observations with a surface detector alone.

Neither the early large surface detector arrays at Volcano Ranch, Haverah Park, SUGAR or Yakutsk, nor the Fly's Eye detector proved big enough to establish the shape of the cosmic-ray spectrum above 10 EeV. Accordingly, second-generation detectors were constructed by the Fly's Eye group (the HiRes instrument) and by a Japanese team who built a detector of 100 km^2. This array, known as AGASA, comprised 111 scintillation detectors each of 2.2 m^2 spaced on a grid of about 1 km spacing. The scintillators were 5 cm thick and so respond mainly to electrons and muons. The detectors were connected and controlled through a sophisticated optical fibre network. The largest events detected have energies of 2 and 3 $\times 10^{20}$ eV and one of these is shown in Fig. 4 [9]. The AGASA array was closed down in January 2004.

Since mid-1998 the HiRes [10] instrument has been taking data at a site in the Dugway desert, near Salt Lake City. This instrument is a stereo system which is used to measure the depth of shower maximum to within 30 g cm^{-2} on an event-by-event basis. This precision was designed to be usefully smaller than the expected difference in the mean depth of maxima for proton or Fe initiated showers. The two locations for the detectors are separated by 12.5 km. The increase in aperture and in $X_{\max}$ resolution over Fly's Eye comes from the reduction in the aperture of each photomultiplier from 5×5 to 1×1 and the increase in the diameter of the mirrors from 1.5 to 2 m. Each mirror is viewed by 256 close-packed photomultipliers: there are 42 mirrors at one site and 22 at the other. The HiRes instrument will cease operation in March 2006.

The energy spectra reported by AGASA [11] (essentially the final version) and by the HiRes group [12] are compared in Fig. 5. It is clear that while the general shape is the same between about 3 and 70 EeV, there are significant differences in intensity at the lower and upper ends of the energy range. In particular, the difference in the number of events claimed above 100 EeV is marked with 11 reported

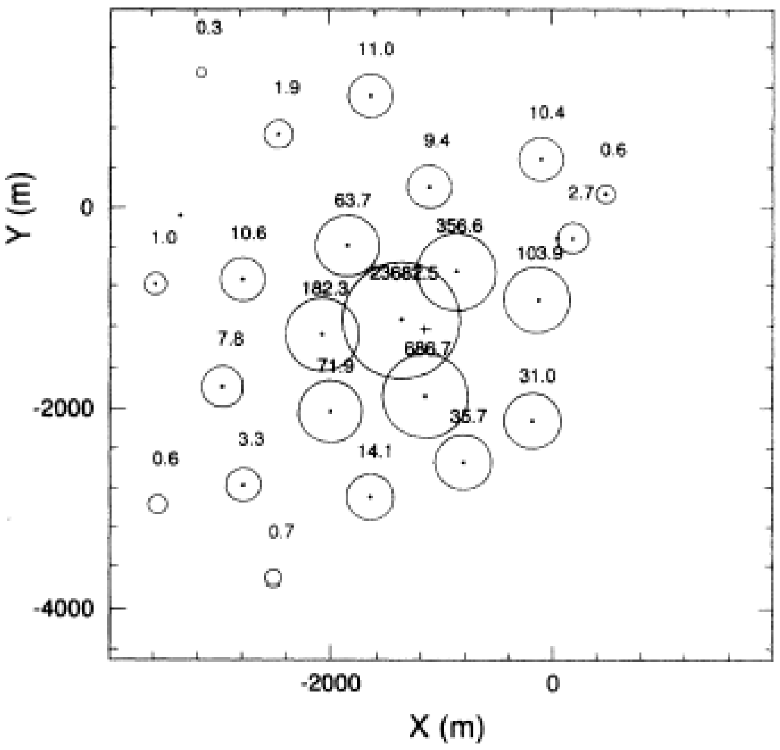


Fig. 4: Map of the detectors struck by the largest event recorded at AGASA. The radius of each circle represent the logarithm of the density recorded at that location. The cross shows the estimated position of the shower core.

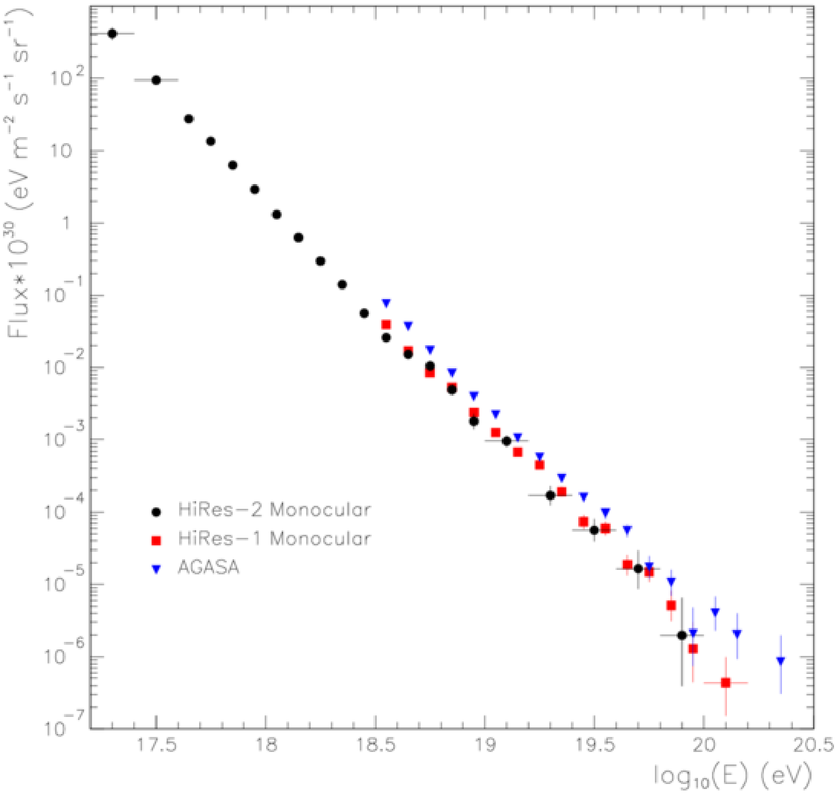


Fig. 5: A comparison of the energy spectra reported by the AGASA and HiRes groups

by AGASA compared with only 2 by HiRes in an exposure that is about three times as great. The reason for these differences is not understood but it is clear that the statistical sample is simply too small.

This problem of low statistics was recognized in the 1980s, even before the AGASA and HiRes detectors had completed construction, and led to the idea that 1000 km^2 of instrumented area was needed if progress was to be made [13]. Jim Cronin argued that 1000 km^2 was insufficiently ambitious and in the summer of 1991 he and Alan Watson decided to try to form a collaboration to build a detector of 5000 km^2, initially without any fluorescence devices. An international workshop [14], organized in Paris by Murat Boratav in 1992, led to a number of focused studies that culminated in a 6-month Design Study during 1995 hosted at the Fermi National Accelerator Laboratory by the then Director, John Peoples.

3 The design of the Auger detector and formation of the collaboration

The philosophy that guided the early phases of the Auger Design Study was to 'let a thousand flowers bloom'. While there was little choice in the form of fluorescence detector that could be used, the possibility of resistive plate counters, scintillators, radio detectors and water-Cherenkov detectors as the devices for the surface array were all discussed and evaluated intensively during the first three months of the workshop. Eventually the choice of water-Cherenkov detectors was made with the intention of having the surface detector (SD) array overlooked by a set of fluorescence detectors (Fig. 6). The water tanks have a significant advantage over scintillators in that their depth (1.2 m of water in the Auger and Haverah Park designs) means that the SD responds to showers coming from a very wide range of zenith angles with relatively high efficiency. Not only does this increase the aperture in a very useful way, but it also opens the possibility of the detection of very high energy neutrinos.

The beauty and power of the Auger Observatory lies in its hybrid nature. The potential of the combination of a SD array and a set of fluorescence detectors (FDs) is still being discovered as it clearly extends the ideas of 1995, but the initial concept of being able to calibrate the 'shower size' recorded with the SD using the 10% or so of events that fall during clear moonless nights has already proved its worth. In addition, the improved geometrical reconstruction available if there is a signal in even one of the 10 m^2 water tanks is very powerful and extends the energy reach of the array to well below 1 EeV. During the Design Study, many aspects of the design and of the projected physics output were explored using an extensive set of Monte Carlo calculations. Detailed simulations of the performance of the ground array for energy and direction measurement were made. At 40 EeV the energy resolution, with the ground array of particle detectors alone, was predicted to be $\sim$10% and the angular resolution $\sim$1.5: on average about 11 detectors were predicted to be struck. The energy resolution and angular accuracy improve as the energy increases. All of these numbers have now been confirmed with real data.

At the design stage the area to be covered at a single site was reduced to 3000 km^2 while it was recognized that with the water-Cherenkov detectors it was possible to envisage all-sky coverage with only two such detectors.

It is one thing to make a design and quite another to find places that might host such large devices. Site surveys were carried out contemporaneously with the Design Study, by Ken Gibbs and Antoine Letessier-Selvon who visited many countries around the world, in both hemispheres. Their brief was to locate places between 1000 and 200 m above sea-level, at a latitude between 30° and 45° north and south of the equator, of 3500 km^2 area and relatively flat with low cloud cover, good visibility, and few local light sources. In addition, access to radio licences, and suitable infrastructure support from national scientists were deemed essential. Argentina was able to offer several potential sites and in November 1995 the decision was made to go there rather than to possible sites in South Africa and Australia. About a year later Millard County in Utah, USA was chosen for the northern site, though the northern location was changed to South Eastern Colorado in 2005.

From the earliest days, a major problem with developing the project was lack of money. Here the influence of Jim Cronin was of immense importance as he was able to get access to people (and

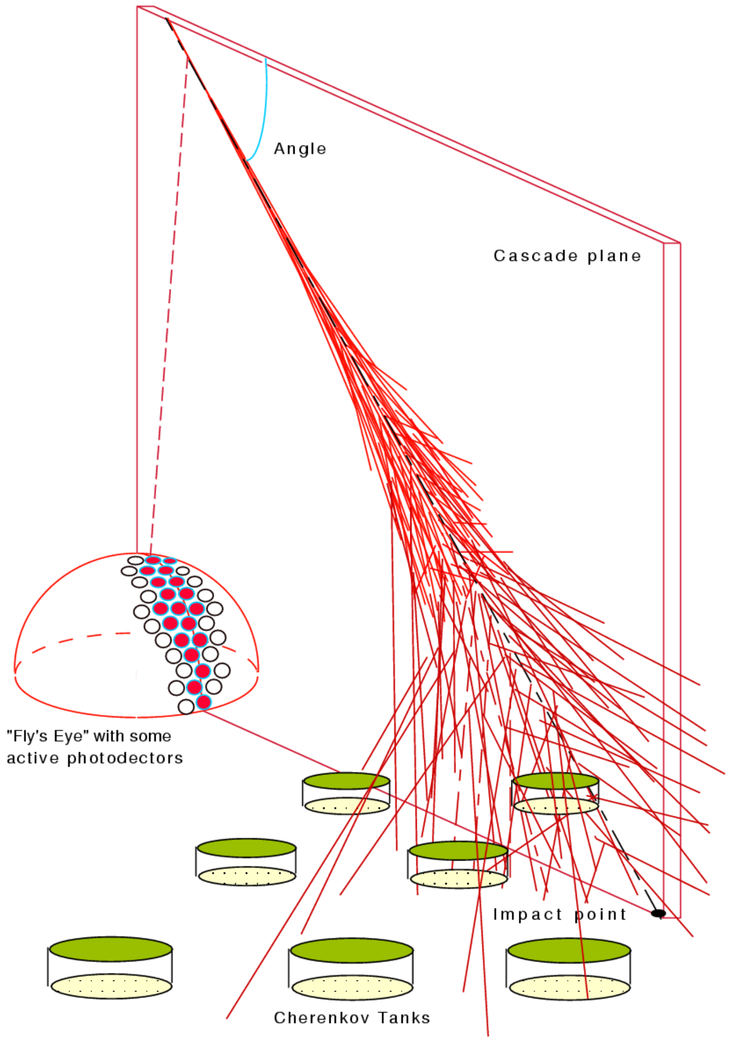


Fig. 6: The conceptual design of the Pierre Auger Observatory. A fluorescence detector overlooks an array of water-Cherenkov detectors. This instructive diagram is due to Enrique Zas (Santiago de Compostela).

extract money from them) when for most of us it would have been difficult to knock on the right door! In particular, money was obtained from the Director-General of UNESCO (although the USA was not a member) and from private individuals whom Jim Cronin knew. The UNESCO money allowed scientists from countries such as Russia, China, and Vietnam to be involved in the design phase. Money for research and development was found from local sources within the laboratories that showed an early interest in the project. For example, at the University of Leeds, lead that had been used for muon shielding and the aluminium lids of the Haverah Park water tanks were sold to help the development of the GPS method that is used to make the relative timing measurements at the detectors [15].

A further problem was the need to fight to have the project recognized as one worthy of attention. Now astroparticle physics is almost an established discipline but this was not so a decade ago and many talks had to be given to raise the awareness of top-class scientists who might be persuaded to join the project and others who might be part of the financial decision-making process. The capital cost was estimated as $100 M for the twosites and forming from scratch the critical mass of competent people that could command this sort of support for a cosmic-ray project was not easy. A particular vulnerability, as with high-energy neutrino astronomy and, to a lesser extent, ground-based gamma-ray astronomy, is that there are no hard theoretical numbers demanding the construction of an instrument of a certain size. This is quite different from the situation with the search for the W and Z, or for the Higgs particle.

Additionally there was the issue of how the project was to be assessed. In particle physics or astrophysics one has become accustomed to umbrella organizations such as CERN, FNAL, ESO and ESA that have developed well-trusted mechanisms over the years for evaluating proposals—no matter how crazy they may seem. We had no such umbrella, so Jim Cronin had the idea of forming our own evaluation panel of scientists of the highest reputation. It was chaired by Professor Ian Axford (a well-known cosmic-ray physicist and then the Director of the Max Planck Institute at Lindau) and included J. Steinberger and M. Koshiba in the committee of six. An extremely favourable report was delivered that was useful in dealing with some agencies, although one agency remarked “of course it was favourable: you chose the panel”!

A major hurdle to overcome was funding from the US. Although no country supplies a majority of the funding, several agencies saw the outcome of debates within the SAGENAP Committee of NSF and DoE as being important input to their own decision making. In the end, in the spring of 1998, the SAGENAP Committee awarded the US groups funding but for only one site and stated that construction should be carried out first in Argentina. By mid-March 1999, a sufficiently strong collaboration from 12 countries (Argentina, Australia, Brazil, Czech Republic, France, Germany, Italy, Mexico, Poland, Slovenia, United Kingdom and the United States) and with sufficient funding to take the first steps had been created: a ground-breaking ceremony took place. Spain joined the Collaboration in 2001 and the Netherlands in 2005. From the beginning, Bolivia and Vietnam were associate members that contribute no funds but students from these countries have the opportunity of training within member countries, thus continuing the spirit behind the early UNESCO support.

The Argentinian site chosen is close to the town of Malargüe (Fig. 7), about five hours by road, south of the city of Mendoza, capital of Mendoza province in western Argentina. The town is well-equipped with hotels and restaurants and a campus site on the edge of the town was made available. It houses an assembly building and office block, designed and built for the project.

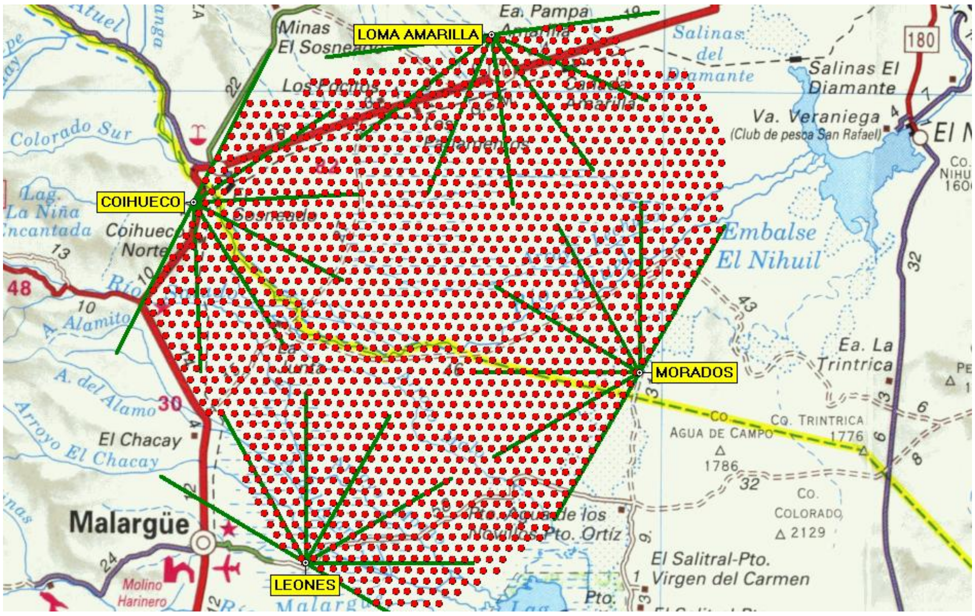


Fig. 7: The planned layout of the Pierre Auger Observatory with 1600 water tanks overlooked by four fluorescence detectors. The water-tank spacing is 1.5 km. As of 31 December 2005 over 1000 surface detectors were taking data, with three of the four fluorescence detectors fully operational. A laser facility, near the centre of the array, is discussed in the text.

4 Early phases of construction and preliminary results

4.1 Characteristics of the detector

The project has associated with it a Finance Board, set up by, and with membership from, the various funding agencies. The Board required that the Collaboration should first construct an engineering array of 40 water tanks and a section of a fluorescence detector. This work was completed in September 2001 and favourably reviewed by an international panel. All of the sub-systems of the Observatory were demonstrated to have achieved or exceeded their specifications. The first 'hybrid' events were recorded in December 2001 when construction of the full instrument commenced.

Fluorescence detectors and water-Cherenkov detectors had been operated before, though not at quite such difficult locations. A new challenge, however, was how to monitor and trigger with 1600 water tanks, distributed over 3000 km^2 (Fig. 7) and each filled with 12 tonnes of water and viewed by three 9" photomultiplier tubes. It is impractical, for reasons of cost and logistics, to connect such an array by cables or optical fibres. Instead each detector was conceived as an autonomous device, as had been done at the SUGAR array [16], but taking advantage of more than 30 years of technological development. The time of each local tank trigger is determined using the GPS technique [15], power is acquired with solar panels and cellular phone technology is used to bring the autonomous signals to the office building where a computer is used to search for trigger signals that are spatially and temporally clustered. When this happens at the level of three stations (currently about 1000 times per day) all of the data associated with the trigger cluster is acquired. The fluorescence detectors use a conventional source of power and their signals are sent to the centre over commercial microwave links. Details of the Engineering Array have been described [17]: the performance of the production instruments does not differ in significant detail.

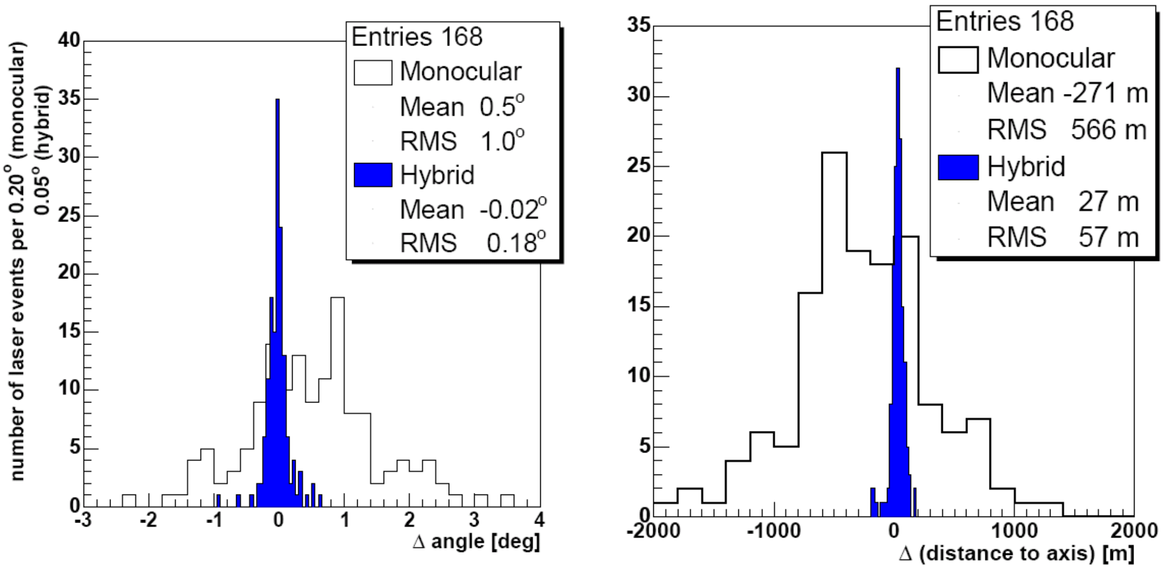
The high level of understanding that is derived from being able to make simultaneous observations of the fluorescence signals and the tank signals is well-illustrated by results from the detection of the scattered light from the Central Laser Facility [18]. This facility, located close to the centre of the array, hosts a 355 nm frequency-tripled YAG laser that generates pulses of up to $\sim$7 mJ. Like the fluorescence detectors, this device is operated remotely from the office building in Malargüe. The scattered light seen at a fluorescence detector from such a pulse is comparable to what is expected from a shower initiated by a primary of 100 EeV. The laser can be pointed in any direction. Some of the light from it is fed into an adjacent tank via an optical fibre so that correlated timing signals can be registered. In this way it has been established that the angular resolution of the surface detectors is $\sim$ 1.7° for $3 < E < 10$ EeV and $\sim$ 0.6° for hybrid events. It has been shown [19] that the accuracy of reconstruction of the position of the laser, using the hybrid technique, is $<$ 60 m. The corresponding figure for the root-mean-square spread, if a monocular reconstruction, is made is $\sim$ 570 m. As there is always at least one tank response in coincidence with each detection at a fluorescence station, these data give a preliminary indication of the geometrical power of the technique. Some results are shown in Fig. 8.

Some idea of the timing accuracy achievable at an individual detector is acquired experimentally from two tanks placed 11 m apart. The data for the twin pair (Carmen and Miranda) of the Engineering Array is shown in Fig. 9. The r.m.s. spread of 23 ns includes the measurement at each detector and the spread in the angles of incidence. After deconvolution, the accuracy is estimated as $\sim$ 12 ns.

The Carmen–Miranda pair is shown in Fig. 10 and in the distant background a fluorescence detector site, Los Leones, is just visible.

In Fig. 11 the 3.5 m × 3.5 m spherical mirror, filter window, and the camera in one of the bays at Los Leones can be seen. Each fluorescence site has six bays. The camera accommodates the 30° azimuth × 28.6° elevation field of view. Each pixel has a field of view of 1.5°.

The data from the surface detectors are displayed on a computer screen at the central office building very shortly after they occur. In Fig. 12 the display for one event is shown. The left-hand part of the top left-hand panel shows the sequence of event triggers with the time and the number of stations triggered



(a) (b)

Fig. 8: (a) Comparison of the angular accuracy achieved by fluorescence detector with and without the benefit of a signal in one water tank; (b) As for Fig. 8(a) but for the distance from the central laser facility.

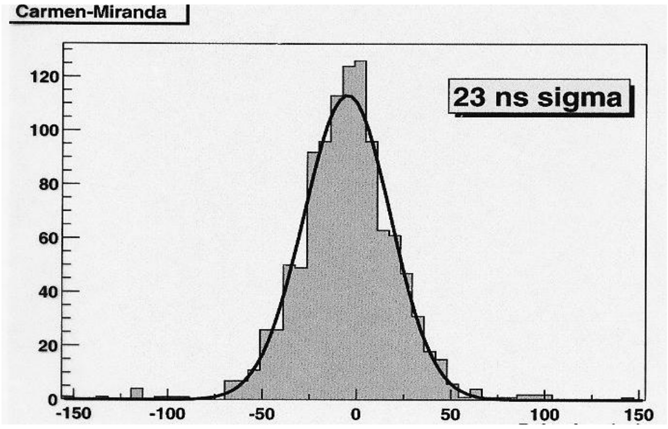


Fig. 9: The timing resolution as deduced from a pair of detectors 11 m apart. The r.m.s. spread of 23 ns includes the measurement at each detector and the spread in the angles of incidence. After deconvolution, the accuracy is estimated as $\sim$ 12 ns.

indicated. Details of the highlighted event are shown in the other panels, including the trigger time and signals at each of the stations (e.g., station 203 had a signal of 625 VEM 3916 ns after the trigger of tank 205). The signal size is measured in units of the signal produced by a 'vertical muon' (VEM). In the event shown fourteen stations have the temporal and spatial characteristics expected and these are displayed in the lower left-hand panel. In the upper-right-hand panel, the fall-off of the signals with distance can be seen. The results of a preliminary analysis are in the lower-right-hand panel.

4.2 Some typical events

In Fig. 13 the signal pattern of a very inclined event (zenith angle = 72°) is shown. The struck detectors are spread out in the azimuthal direction of arrival of the event. The event is about 15 km long and about 5 km wide. Estimating the primary energy of the particles that initiated events such as this is not

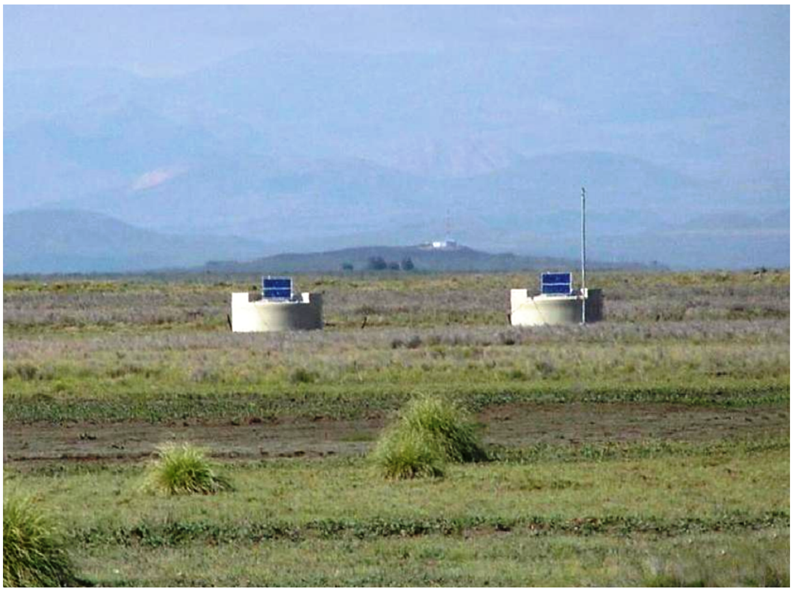
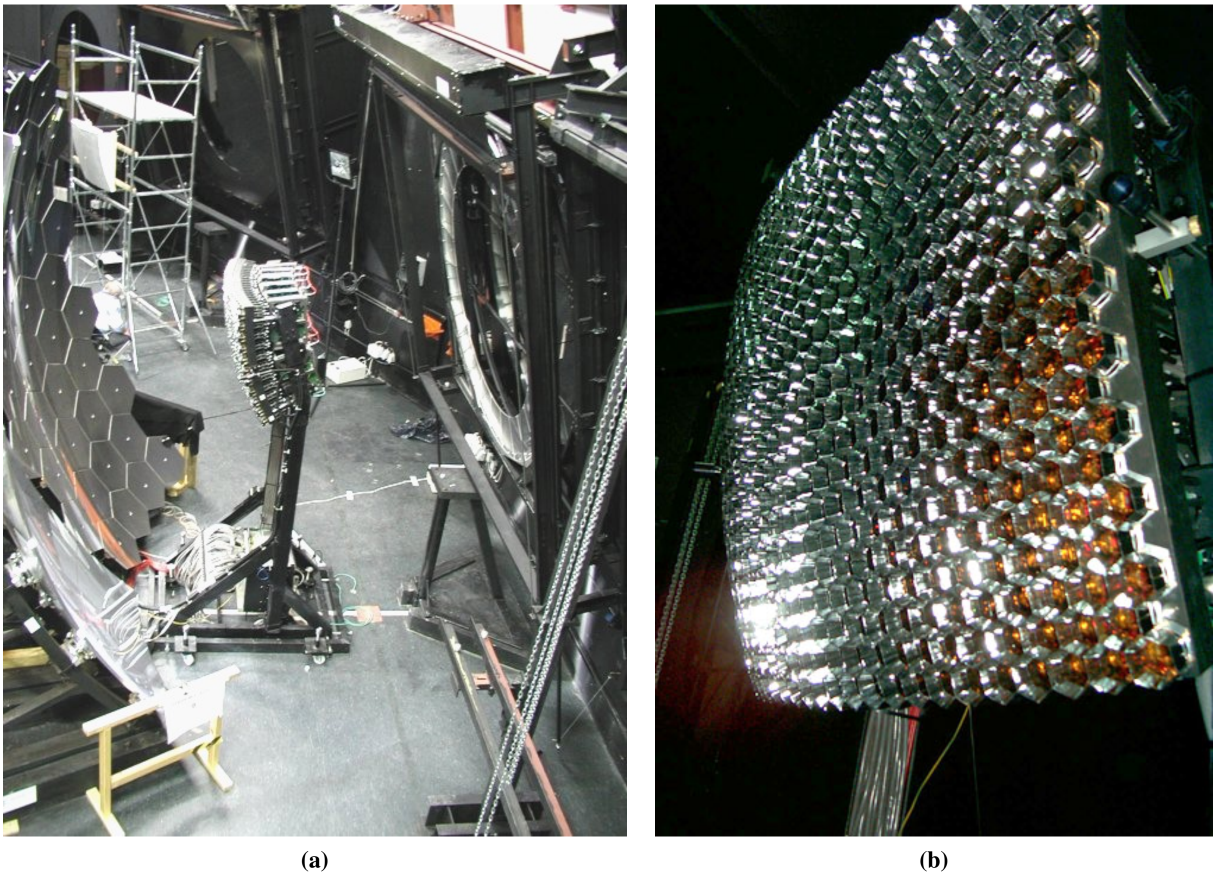


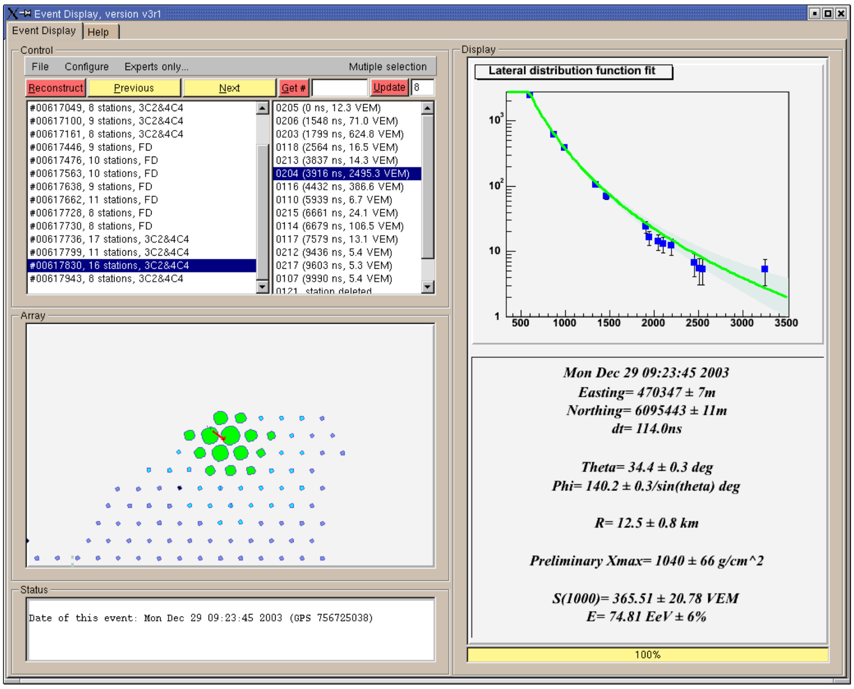
Fig. 10: The detector pair of the Engineering Array, Carmen and Miranda, used for estimating the timing accuracy with which signals are recorded at the detectors. Signal size accuracy is also determined from such pairs, of which there are two on the production array. The Los Leones fluorescence site is visible in the distance between the detectors.



(a)

(b)

Fig. 11: (a) A view of the 3.5 m $\times$ 3.5 m spherical mirror (left) and the aperture/filter through which light is received; (b) One of the 24 cameras used to photograph the fluorescence light. The camera mount can be seen in Fig. 11(a). There are 440 photomultipliers in each camera.



```
#00617049, 8 stations, 3C2&4C4
#00617100, 9 stations, 3C2&4C4
#00617161, 8 stations, 3C2&4C4
#00617446, 9 stations, FD
#00617476, 10 stations, FD
#00617563, 10 stations, FD
#00617638, 9 stations, FD
#00617662, 11 stations, FD
#00617728, 8 stations, FD
#00617730, 8 stations, FD
#00617736, 17 stations, 3C2&4C4
#00617799, 11 stations, 3C2&4C4
#00617830, 16 stations, 3C2&4C4
#00617943, 8 stations, 3C2&4C4
```

```
0205 (0 ns, 12.3 VEM)
0206 (1548 ns, 71.0 VEM)
0203 (1799 ns, 624.8 VEM)
0118 (2564 ns, 16.5 VEM)
0213 (3837 ns, 14.3 VEM)
0204 (3916 ns, 2495.3 VEM)
0116 (4432 ns, 386.6 VEM)
0110 (5939 ns, 6.7 VEM)
0215 (6661 ns, 24.1 VEM)
0114 (6679 ns, 106.5 VEM)
0117 (7579 ns, 13.1 VEM)
0212 (9436 ns, 5.4 VEM)
0217 (9603 ns, 5.3 VEM)
0107 (9990 ns, 5.4 VEM)
0121  station deleted
```

```
Date of this event: Mon Dec 29 09:23:45 2003 (GPS 756725038)
```

Mon Dec 29 09:23:45 2003
Easting= 470347 ± 7m
Northing= 6095443 ± 11m
dt= 114.0ns

Theta= 34.4 ± 0.3 deg
Phi= 140.2 ± 0.3/sin(theta) deg

R= 12.5 ± 0.8 km

Preliminary Xmax= 1040 ± 66 g/cm^2

S(1000)= 365.51 ± 20.78 VEM
E= 74.81 EeV ± 6%

Fig. 12: An example of an event with energy above 10 EeV at 34° from the zenith. Fourteen stations have been struck (see bottom left) and the fall-off of the signal size with distance (the lateral distribution function) shown in the upper-right-hand corner is consistent with expectation. The shower data in the bottom-right-hand panel are taken from the real-time analysis facility and are very preliminary.

straightforward as the shower loses the near-circular symmetry of smaller angles because of bending of the muons (the dominant surviving particles) by the geomagnetic field.

There is great interest in studying inclined events as they may offer a route to the detection of very high energy neutrinos. This idea, first proposed by Berezinsky and Smirnov [20], was re-examined in the context of the Auger Observatory by Capelle *et al.* [21]. The trick is to study the properties of showers that arise at very large angles (>70°) from the vertical (see Fig. 14). A neutrino can interact anywhere in the atmosphere with equal probability. However, if one restricts a search to large zenith angles then it should be possible to identify occasions when the neutrino has interacted deep in the atmosphere. The mode of identification depends on the detection technique. A neutrino-induced shower arriving at a large zenith angle has distinctive characteristics that make it possible to envisage detecting it with a conventional, ground-based, air shower array. Most showers detected at large zenith angles will have been produced by nucleonic primaries. The vast majority of the particles detected in such events will be high-energy muons as at >70° the large atmospheric thickness of more than $\sim 2500\ \mathrm{g\,cm^{-2}}$ (at the depth of the Auger Observatory) filters out the electromagnetic radiation that arises from neutral pion decay. The muons are accompanied by a small fraction of electromagnetic component (around 20%) that is in time and spatial equilibrium with the muons. This electromagnetic component has its origin in muon bremsstrahlung, pair production, knock-on electrons, and muon decay. These showers have large radii of curvature as the source of the muons is far from the shower detector. The particles in the shower disc arrive tightly bunched in time and the distribution of the signal size is rather flat across the array. By contrast, a shower produced by a neutrino, if it interacts in the volume of air over the detector, will have a curved shower front, a steep fall-off of particle signal with distance from the shower core and a distinctively broad time spread of the particles at the detectors.

The only instrument which is currently large enough to have any prospect of detecting neutrinos, and with the ability to exploit these characteristics, is the Pierre Auger Observatory.

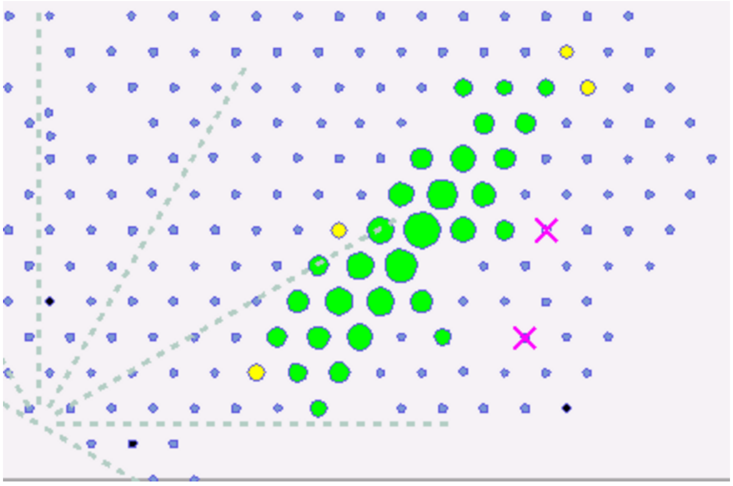
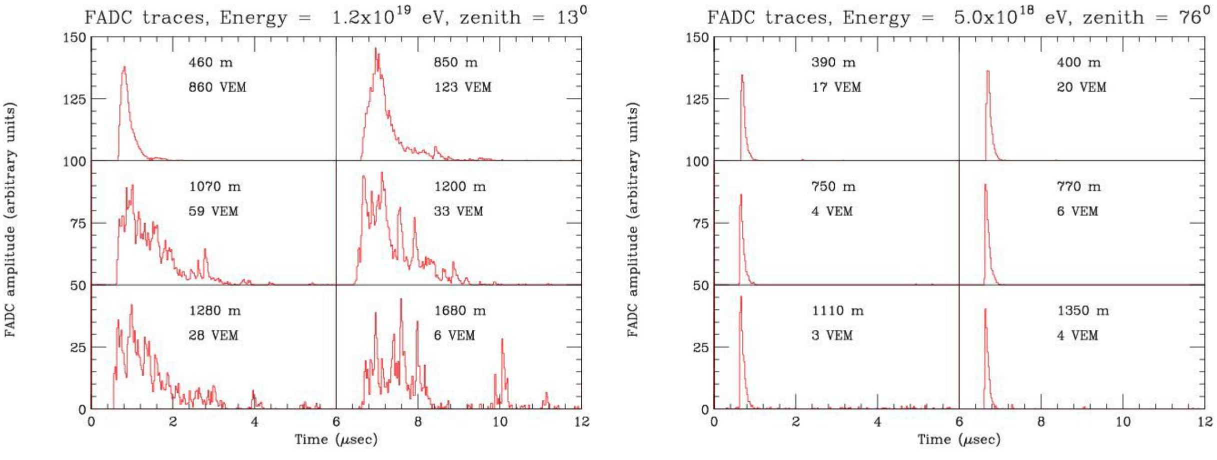


Fig. 13: The density pattern in an inclined event at 72° from the vertical. Thirty-three detectors have been triggered. Those marked with a cross are chance coincidences within the trigger window and are not part of the event. Estimates of the energy of events such as this is made complicated because of the deflections of the constituent muons in the geomagnetic field.



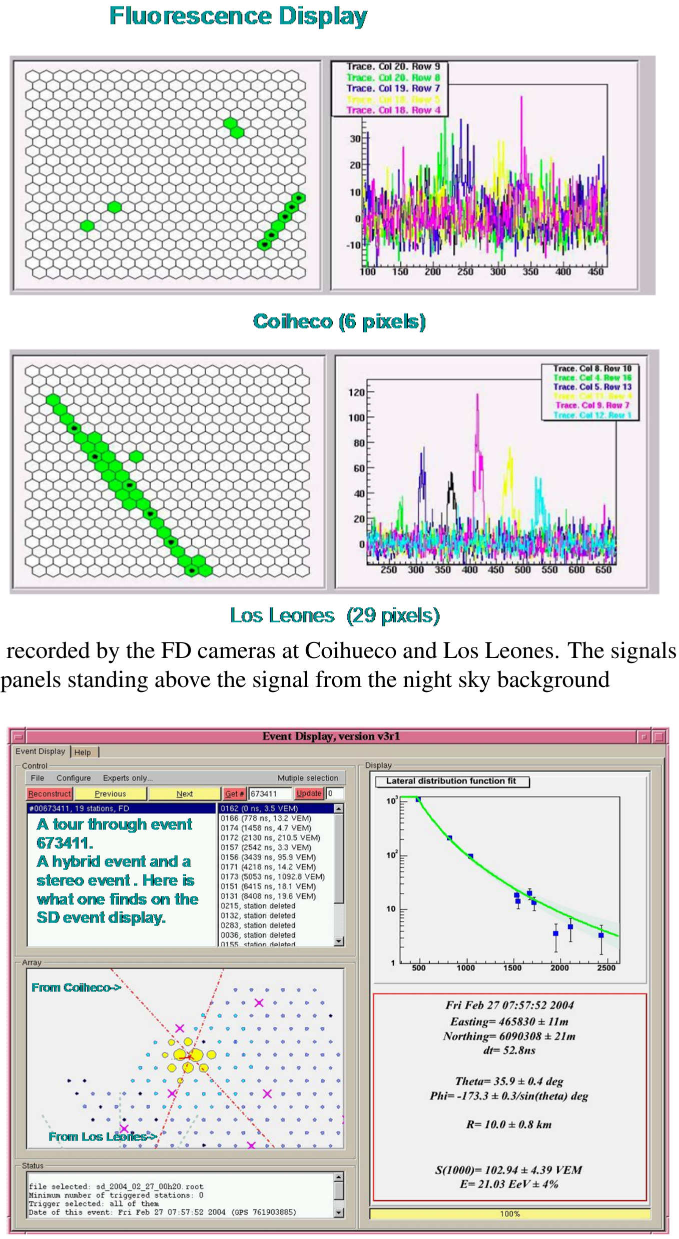
(a) (b)

Fig. 14: (a) The FADC signals from a nearly vertical air-shower. It is evident that the signals become broader as the distance from the shower axis (shown in each panel) increases. The gradient in signal size is also evident (compare the detail in the top right-hand panel in Fig. 12). (b) The FADC signals in an inclined shower. By contrast with Fig. 14(a), the time spread is very small and nearly independent of axial distance. A shower with the characteristics of Fig. 14(a) but at a zenith angle above 70° might well be produced by a neutrino.

An example of a hybrid event is shown in Fig. 15. Figure 15(a) serves also as an example of the type of data that comes from a fluorescence detector. The signals are clearly visible above the night sky background.

Figure 15(b) shows the event display for the surface detectors (compare Fig. 12).

The improvement in the geometrical reconstruction in a hybrid event is shown in Fig. 15(c) (compare Fig. 8 where data from the central laser facility were used).



(a) The fluorescence signals recorded by the FD cameras at Coihueco and Los Leones. The signals from the highlighted pixels are shown in the right-hand panels standing above the signal from the night sky background

#00673411, 19 stations, FD

A tour through event 673411.
A hybrid event and a stereo event . Here is what one finds on the SD event display.

```
0162 (0 ns, 3.5 VEM)
0166 (778 ns, 13.2 VEM)
0174 (1458 ns, 4.7 VEM)
0172 (2130 ns, 210.5 VEM)
0157 (2542 ns, 3.3 VEM)
0156 (3439 ns, 95.9 VEM)
0171 (4218 ns, 14.2 VEM)
0173 (5053 ns, 1092.8 VEM)
0151 (6415 ns, 18.1 VEM)
0131 (8408 ns, 19.6 VEM)
0215, station deleted
0132, station deleted
0283, station deleted
0036, station deleted
0155, station deleted
```

```
file selected: sd_2004_02_27_00h20.root
Minimum number of triggered stations: 0
Trigger selected: all of them
Date of this event: Fri Feb 27 07:57:52 2004 (GPS 761903885)
```

Fri Feb 27 07:57:52 2004
Easting= 465830 ± 11m
Northing= 6090308 ± 21m
dt= 52.8ns

Theta= 35.9 ± 0.4 deg
Phi= -173.3 ± 0.3/sin(theta) deg

R= 10.0 ± 0.8 km

S(1000)= 102.94 ± 4.39 VEM
E= 21.03 EeV ± 4%

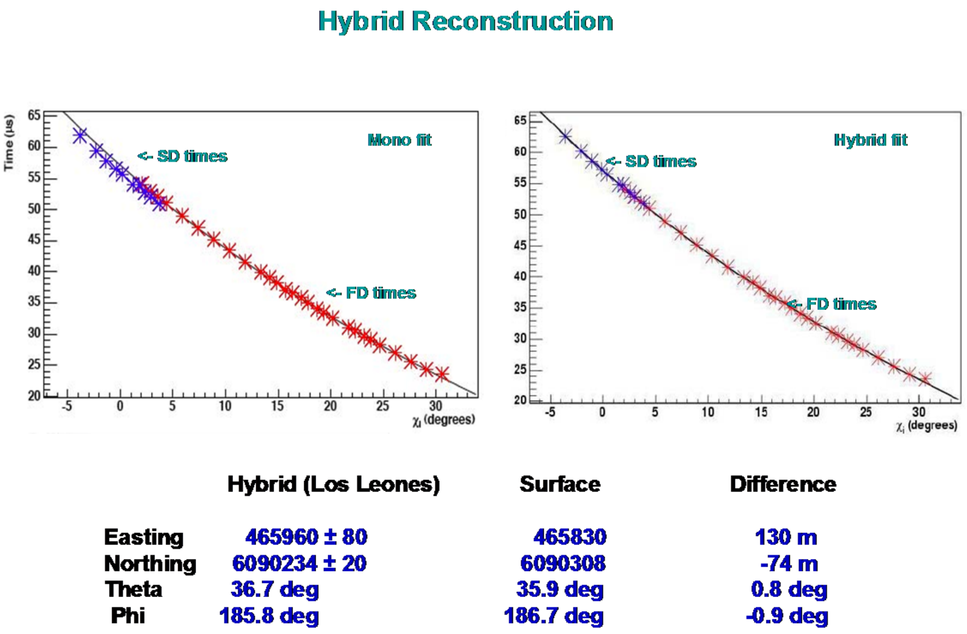
(b) The event display for the SD signals in event 673411 for which the FD signals are shown in Fig. 15(a). The dotted lines in the lower left-hand panel indicate the shower planes derived from the FD signals.

Fig. 15: The hybrid event 673411

4.3 The primary energy spectrum

The Auger Collaboration has reported [22] the first precision measurement of the high-energy cosmic ray spectrum made from the Southern Hemisphere.

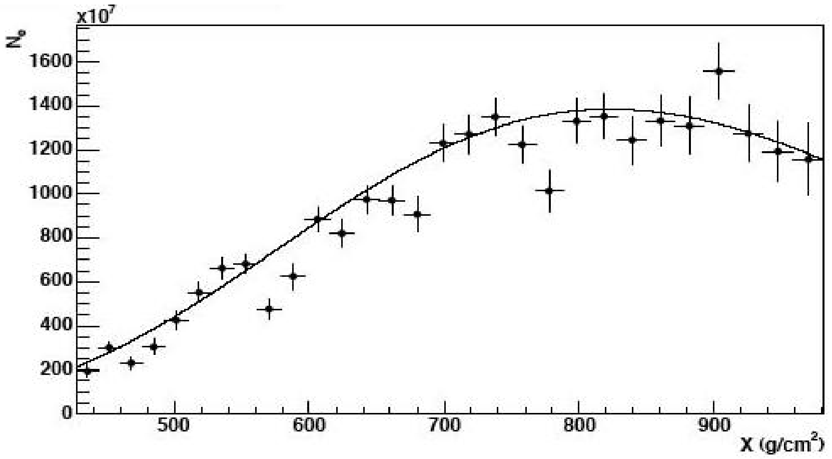
For this analysis attention was restricted to events with zenith angle $\theta < 60^{\circ}$. The strategy was to reconstruct the arrival direction for each event recorded by the SD and to estimate the magnitude of the signal at 1 km from the shower axis, S(1000), as a measure of the size of the shower in units defined by the signal from a muon that traverses the tank vertically. The shower axis S(1000) is chosen as the ground-parameter as it can be measured to better than 10%. In addition, as shown in the pioneering studies of Hillas [23], the size of this ground-parameter is ~ 3 times less susceptible to stochastic fluctuations and variations in primary mass than are measurements made close to the shower axis.



Hybrid Reconstruction

	Hybrid (Los Leones)	Surface	Difference
Easting	465960 ± 80	465830	130 m
Northing	6090234 ± 20	6090308	-74 m
Theta	36.7 deg	35.9 deg	0.8 deg
Phi	185.8 deg	186.7 deg	-0.9 deg

(c) This diagram illustrates the power of using the times at which the shower particles trigger the surface detectors. The left-hand plot shows the poor agreement between the FD times and the SD times from a monocular reconstruction. The combination is shown in the right-hand panel. The difference in distance and angular reconstruction is shown in the table below the panels.



(d) The longitudinal development curve deduced from the fluorescence data. Estimates of the primary energy can be obtained from the number of particles at shower maximum and by integrating under the curve.

Fig. 15: The hybrid event 673411 (*cont.*)

Two cosmic rays of the same energy, but incident at different zenith angles, will yield different values of S(1000). Thus a necessary step is to find the relation between the ground-parameter measured at one zenith angle and that measured at another. The approach adopted here is to use the well-established technique of the constant intensity cut (CIC) method which has been recently reappraised [24]. The principle of this method is that the high level of isotropy of cosmic rays leads to the proposition that showers created by primaries of the same mass and energy will be detected at the observation level at the same rate. Here the rate of events above different S(1000) is found for different zenith angles and all azimuth angles so that events come from a broad band of sky. This method is used to establish the relationship between $S(1000)_{38^\circ}$ and $S(1000)_\theta$, where the subscripts refer to a reference angle, chosen as 38°, and θ is the angle of incidence. The average thickness of the atmosphere above the Auger Observatory is 875.5 $\mathrm{g\,cm^{-2}}$.

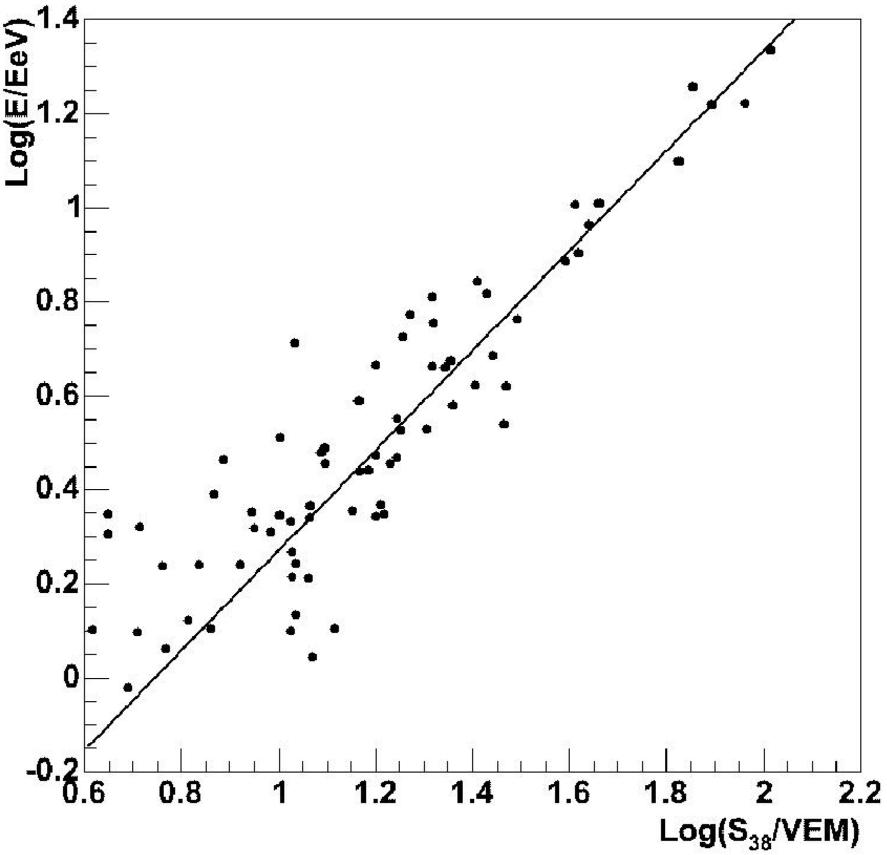
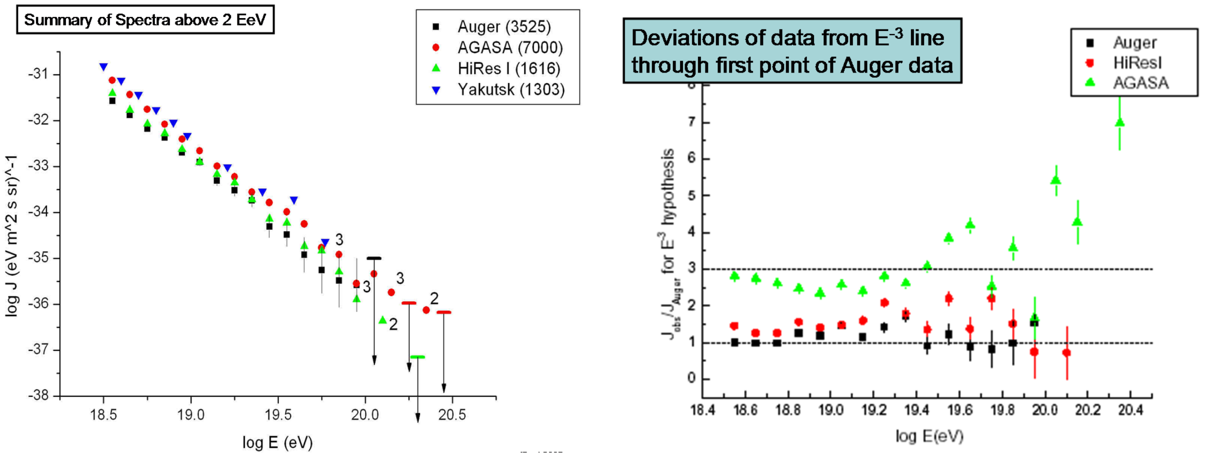


Fig. 16: A comparison of the energy estimated from the fluorescence detectors with the signal size (normalized to 38°) observed at 1000 m from the axis of the shower. The FD energy has been corrected by ∼10% for the missing energy carried into the ground by high-energy muons and neutrinos.

The link between S(1000)$_{38^\circ}$ and the primary energy is established using data from the fluorescence detectors rather than through model calculations. On clear, moonless nights, it is possible to observe fluorescence signals simultaneously with the SD events: this 'hybrid' approach, a key characteristic of the Auger Observatory, offers several advantages. For every FD event for which the shower core falls within the instrumented SD area, at least one tank is struck so that the time at which the tank was triggered can be used to enhance the reconstruction of the FD geometry. Further, as the FD instruments are used primarily as calibration devices in this application, the selection of events can be made in a highly selective manner. This was done in Ref. [22], where the FD tracks had to be longer than 350 $\mathrm{g\,cm^{-2}}$, the contribution of the Cherenkov light to the signals collected less than 10%, and there were contemporaneous measurements of the aerosol content of the atmosphere, as was possible in the latter part of the data run. There are significant systematic uncertainties currently present in the Auger spectrum arising largely from the lack of knowledge of the fluorescence yield of atmospheric nitrogen and from the low statistics available for the S(1000)$_{38^\circ}$ energy calibration. At 3 EeV the systematic uncertainty is about 30% growing to 50% at 100 EeV.

When estimating the energy of an event from the fluorescence yield (Fig. 16), a correction must be made for 'missing energy' carried by high-energy muons and neutrinos. A study of this conversion factor has recently been made for nucleonic primaries with a variety of hadronic interaction models. At 10 EeV the correction for missing energy is ∼10% with a systematic uncertainty, due to our lack of knowledge of the nuclear mass and the hadronic interactions, estimated as ∼7% [25]. The corrections and the associated systematic uncertainties may have to be revised when LHC data are available.



(a) The differential spectra from Auger, AGASA, HiResI and Yakutsk are compared on a plot of log J vs. log E. The numbers shown in the legend correspond to the events reported above 3 EeV. The numbers (3, 2) by some points refer to the last bin of each data set in which > 0 events were recorded.

(b) The ratio of the values of each point with respect to a fit of E^{-3} to the first point of the Auger spectrum at 3.55 EeV which contains 1216 events. The purpose of the plot is to illustrate the differences between the different measurements in a straightforward manner. Yakutsk data are not included in this plot as they are so discordant.

Fig. 17: Experimental spectra obtained by different groups

A comparison of the spectra reported by the different groups is made in Fig. 17. The agreement is poor, even at 10 EeV where there may still be differences of ~ 2 between the fluxes from the different instruments.

There is clearly scope for much further work on analysis and on understanding hadronic interactions and perhaps some of this will appeal to students of the School.

Acknowledgements

I would like to thank the organizers of the CERN–CLAF School of 2005 for the invitation to give a talk about the Pierre Auger Observatory. It was a pleasure to spend time with such bright and interesting students. The Auger Observatory would not have reached its present stage without the efforts of many people but I would particularly like to mention the efforts of Murat Boratav, Jim Cronin and Paul Mantsch. I am most grateful to Joerg Hoerandel for making me aware of Köhlhorster's work cited in Ref. [3]. Research on ultra-high-energy cosmic rays at the University of Leeds is supported by the Particle Physics and Astronomy Research Council, UK.

References

[1] P. Auger *et al., Comptes Rendus* **206** (1938) 721; P. Auger *et al., Rev. Mod. Phys.* **11** (1939) 288.

[2] B. Rossi, *Supplemento a La Ricerca Scientifica* **1** (1934) 579.

[3] W. Kolhörster *et al., Naturwissenschaften* **26** (1938) 576.

[4] J. Linsley, *Phys. Rev. Lett.* **10** (1963) 146.

[5] K. Greisen, *Phys. Rev. Lett.* **16** (1966) 748; G.T. Zatsepin and V.A. Kuz'min, *Zh Eksp. Teor. Fiz. Pis'ma Red* **4** (1966) 144.

[6] T.K. Gaisser, from talk at Leeds Symposium, July 2004.

[7] K. Eggert, *Nucl. Phys. B (Proc. Suppl.)* **122** (2003) 447.

[8] D. Bird *et al., Astrophys. J.* **441** (1995) 144.

[9] M. Hayashida *et al., Phys. Rev. Lett.* **73** (1994) 3491.

[10] P. Sokolsky, *AIP Conference Proceedings,* **433** (1998), p. 65.

[11] M. Takeda *et al., Astropart. Phys.* **18** (2003) 135.

[12] R.U. Abbassi, *Phys. Lett.* **B619** (2005) 271.

[13] A.A. Watson, *Nucl. Phys. B (Proc. Suppl.)* **22B** (1991) 116.

[14] Proceedings of the International Workshop on Techniques to Study Cosmic Rays with Energies $>$ 10^{19} eV, *Nucl. Phys. B (Proc. Suppl.)* **28B** (1992), Eds. M. Boratav, J.W. Cronin and A.A. Watson; J.W. Cronin, *Nucl. Phys. B (Proc. Suppl.)* **28B** (1992) 213.

[15] C.L. Pryke and J. Lloyd-Evans, *Nucl. Instrum. Methods* **A354** (1995) 560.

[16] M.M. Winn *et al., J. Phys. G* **12** (1986) 675.

[17] J. Abraham *et al.* [Auger Collaboration], *Nucl. Instrum. Methods* **A523** (2004) 50.

[18] F Arqueros *et al.* (Pierre Auger Collaboration), *Proc. 29th ICRC*, Pune, India, 2005 (Tata Institute of Fundamental Research, Mumbai, India, 2005), vol. 8, pp. 335–338. `http://icrc2005.tifr.res.in/htm/Vol-Web/Vol-18/18335-usa-malek-M-abs1-he15-poster.pdf`

[19] M.A. Mostafá (Pierre Auger Collaboration), *Proc. 29th ICRC 2005*, Pune, India, 2005 (Tata Institute of Fundamental Research, Mumbai, India, 2005), vol. 7, pp. 369–372. `http://icrc2005.tifr.res.in/htm/Vol-Web/Vol-17/17369-usa-mostafa-M-abs1-he14-oral.pdf`

[20] V.S. Berezinsky and A.Yu. Smirnov, *Astrophys. Space Sci.* **32** (1975) 461.

[21] K.S. Capelle *et al., Astropart. Phys.* **8** (1998) 321.

[22] P. Sommers (Pierre Auger Collaboration), *Proc. 29th ICRC*, Pune, India, 2005 (Tata Institute of Fundamental Research, Mumbai, India, 2005), vol. 7, pp. 387–390 [astro-ph/0507150]. `http://icrc2005.tifr.res.in/htm/Vol-Web/Vol-17/17387-usa-sommers-P-abs1-he14-oral.pdf`

[23] A.M. Hillas, *Acta Phys. Acad. Sci. Hung.* **29** *Suppl. 3* (1969) 355; A.M. Hillas *et al., Proc. 12th ICRC*, Hobart, Australia, **3** (1971) 1001.

[24] J. Alvarez-Muñiz *et al., Phys. Rev.* **D66** (2002) 123004 [astro-ph/0209117].

[25] T. Pierog *et al.*, (Pierre Auger Collaboration), *Proc. 29th ICRC*, Pune, India, 2005 (Tata Institute of Fundamental Research, Mumbai, India, 2005), vol. 7, pp. 103–106. `http://icrc2005.tifr.res.in/htm/Vol-Web/Vol-17/17103-ger-engel-R-abs2-he14-oral.pdf`

Instrumentation for high-energy physics

S. Stapnes
University of Oslo, Norway

Abstract

The first part of this summary contains a description of the passage of particles through matter. The basic physics processes for charged particles, photons, neutrons and neutrinos are mostly electromagnetic (collision losses described by Bethe–Bloch, bremsstrahlung, photo-electric effect, Compton scattering and pair production) for charged particles and photons; additional strong interactions for hadrons; neutrinos interacting weakly with matter. Concepts like radiation length, electromagnetic showers, nuclear interaction/absorption length and showers are covered. Important processes like multiple scattering, Cherenkov radiation, transition radiation, and dE/dx for particle identification are described next. This is followed by a short discussion of momentum measurement in magnetic fields. The last part of the summary covers particle detection by means of ionization detectors, scintillation detectors and semiconductor detectors. Signal processing is briefly discussed at the end.

1 Introduction

Experimental particle physics is based on many advanced instruments and methods. The main instruments are accelerators, with key parameters such as luminosity, energy and particle type. Next follow the detectors, whose key parameters are efficiency, speed, granularity and resolution. The online data-processing and the trigger/DAQ have to operate with high efficiency, large compression factors and throughput, and be optimized for a number of physics channels. The offline analysis aims to extract and understand signal and background and ultimately improve our physics models and understanding. In this chain we should keep in mind that the primary factors for a successful physics measurement are the accelerator and detector/trigger systems and that losses there are not recoverable. New and improved detectors are therefore extremely important for our field.

2 Energy loss in matter

We shall first concentrate on electromagnetic forces since a combination of their strength and range make them the primary cause of energy loss in matter. For neutrons, hadrons generally and neutrinos, strong and weak interactions also have an effect.

A unified approach to the energy loss of a charged particle in matter due to electromagnetic forces can be found in Ref. [1] and its references. By considering the electromagnetic interaction between a charged particle with a certain mass and velocity, and a material with a given refractive index and dielectric constant, a photon with a certain energy can be created. At photon energies below the excitation energy of the material, Cherenkov light is created if the velocity of the particle in the material is greater than the velocity of light in the medium. At slightly higher photon energies, virtual photons are exchanged between the incoming particle and the atoms, resulting in excitations and ionizations. Finally, X-ray photons (transition radiation) can be emitted if there are discontinuities in the material traversed by the particle. In the following, these three effects are discussed separately.

2.1 Heavy charged particles

Heavy charged particles transfer energy mostly to the atomic electrons causing ionization and excitation. We shall come back later to light charged particles, in particular electrons and positrons. Usually the

Bethe–Bloch formula is used to describe the energy loss of heavy charged particles. Most of the features of the formula can be understood from a very simple model.

1. Consider the energy transfer to a single electron from a heavy charged particle passing at a distance b.
2. Multiply by the number of electrons passed.
3. Integrate over all reasonable distances b.

The impulse transferred to the electron will be

$$I = \int F\, dt = e \int E_{\perp} \frac{dx}{\nu} = \frac{2ze^2}{b\nu}\,.$$

The integral is solved by using Gauss's law over an infinite cylinder centred along the particle track. The energy transfer is therefore

$$\Delta E(b) = \frac{I^2}{2m_e}\,.$$

The energy transfer to a volume dV, where the electron density is N_e, can now be calculated:

$$-dE(b) = \Delta E(b)\, N_e\, dV_e\,; \quad dV = 2\pi b\, db\, dx\,.$$

The energy loss per unit length is hence given by

$$-\frac{dE}{dx} = \frac{4\pi z^2 e^4}{m_e \nu^2}\, N_e\, \ln \frac{b_{\text{max}}}{b_{\text{min}}}\,.$$

The distance b_{min} is not zero but can be determined by the maximum energy transferred in a head-on collision. The distance b_{max} is given by the requirement that the perturbation be short relative to the period $(1/\nu)$ of the electron.

We end up with the following:

$$-\frac{dE}{dx} = \frac{4\pi z^2 e^4}{m_e \nu^2}\, N_e\, \ln \frac{\gamma^2 m_e \nu^3}{z e^2 \bar{\nu}}\,,$$

which should be compared to the Bethe–Bloch formula below. (Note: dx in Bethe–Bloch includes density, so the unit is g cm^{-2}.)

$$\langle\frac{dE}{dx}\rangle = -4\pi N_A\, r_e^2\, m_e\, c^2\, z^2\, \frac{Z}{A}\, \frac{1}{\beta^2} \left[\frac{1}{2} \ln \frac{2 m_e c^2 \gamma^2 \beta^2}{I^2}\, T_{\text{max}} - \beta^2 - \frac{\delta}{2}\right]\,.$$

Bethe–Bloch parametrizes over momentum transfers using I (the ionization potential) and T_{max} (the maximum transferred in a single collision). The correction δ describes the effect that the electric field of the particle tends to polarize the atoms along its path, hence protecting electrons far away (this leads to a reduction/plateau at high energies). The curve has a minimum at β = 0.96 ($\gamma\beta$ = 3.5) and increases slightly for higher energies; for most practical purposes, one can say that the curve depends only on β (in a given material). Below the minimum ionizing point, the curve follows $\beta^{-5/3}$. At low energies, other models are useful (as shown in Fig. 1).

The radiative losses seen in Fig. 1 at high energy will be discussed later (in connection with electrons where they are much more significant at lower energies).

Since particles with different masses have different momentum for the same β, the dE/dx curves for protons, pions, kaons, etc. are shifted with respect to each other along the x-axis when dE/dx is plotted as a function of momentum. This can be used for particle identification at relatively low energies in tracking chambers (see Section 3.3).

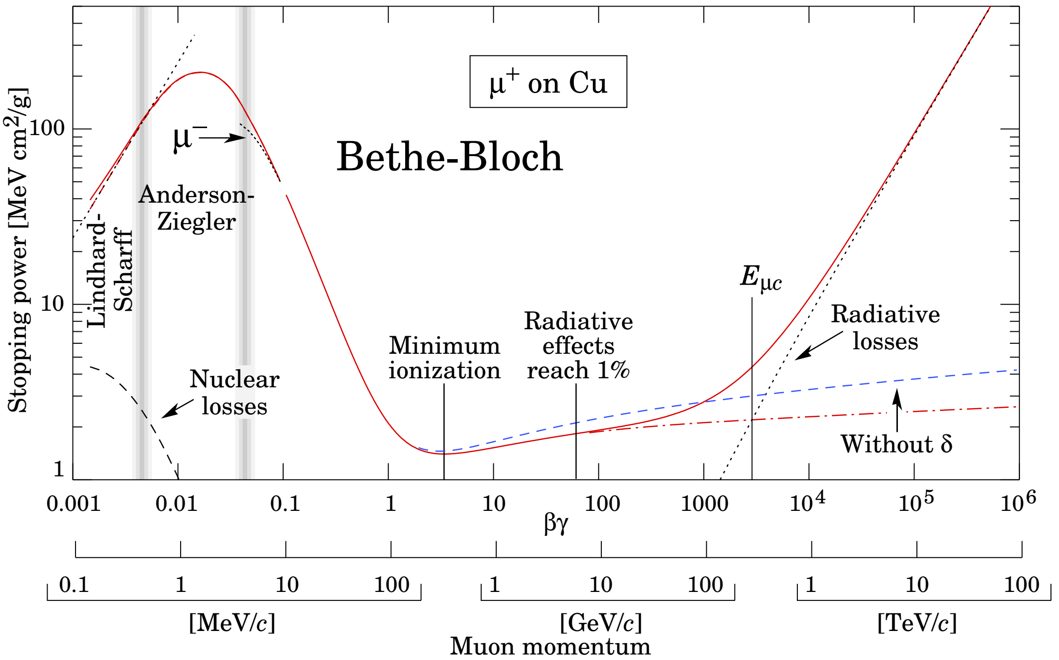


Fig. 1: Radiation loss of muons in matter. From Ref. [3].

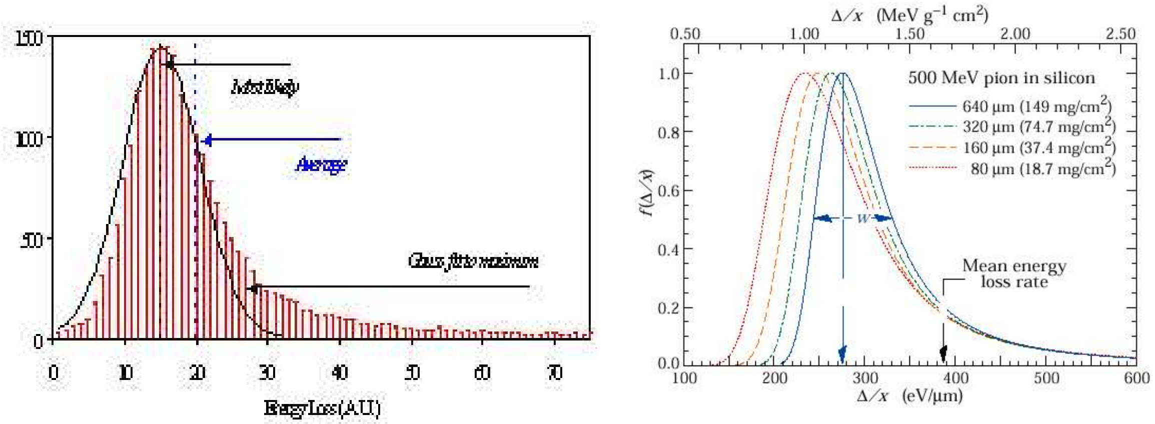


Fig. 2: Distribution of energy loss in absorbers of varying thickness. From Refs. [2, 3].

While Bethe–Bloch describes the average energy deposition, the probability distribution in thin absorbers is described by a Landau distribution. Other functions are often used: Vavilov for slightly thicker absorbers, Bischel, etc. [2, 3].

In general, these are skewed distributions (Fig. 2) tending towards a Gaussian when the energy loss becomes large (in thick absorbers). One can use the ratio between energy loss in the absorber under study and T_{max} from Bethe–Bloch to characterize the absorber thickness.

2.2 Light charged particles: electrons and positrons

For electrons and positrons the Bethe–Bloch formula has to be modified to take into account that the incoming particle has the same mass as the atomic electrons. In addition, a significant amount of energy is carried away by bremsstrahlung photons. The cross-section for this process goes as $1/m^2$ and is therefore very significant for electrons and positrons even though it also plays a role at higher energy for muons, as seen in Fig. 1. The differential cross-section for bremsstrahlung (ν is the photon frequency)

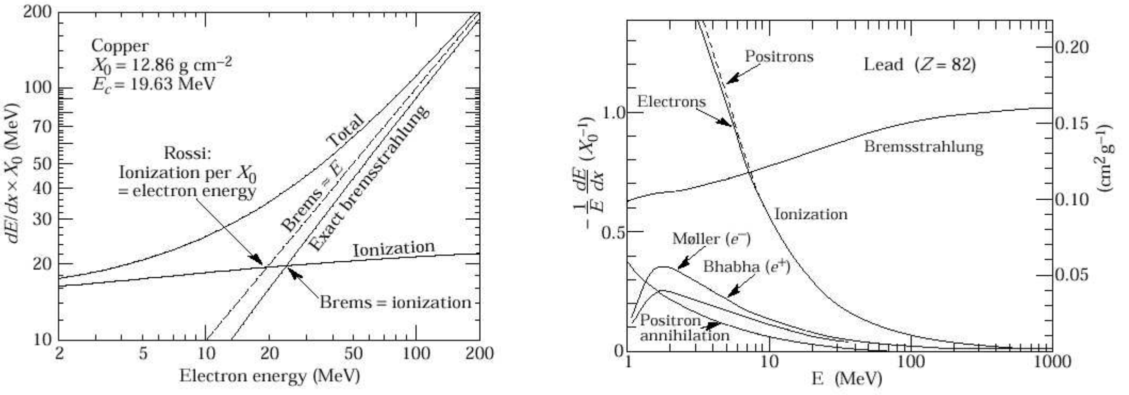


Fig. 3: Energy loss of electrons in copper and lead as a function of electron energy. The critical energy E_C is defined as the point where the ionization loss is equal to the bremsstrahlung loss.

in the electric field of a nucleus with atomic number Z is given approximately by

$$d\sigma \propto Z^2 \, \frac{d\nu}{\nu} \, .$$

The bremsstrahlung loss is therefore

$$-\left(\frac{dE}{dx}\right) = N \int_0^{\nu_0 = E_0/h} h\,\nu \, \frac{d\sigma}{d\nu} \, d\nu = N \, E_0 \, \Phi(Z^2) \, ,$$

where the linear dependence on energy is apparent. The Φ function depends mostly on the material; for example, on the square of the atomic number as shown. Here N is the atom density of the material. Bremsstrahlung in the field of the atomic electrons must be added (giving $Z^2 + Z$). The above equation can be rewritten as

$$-\left(\frac{dE}{E}\right) = N \, \Phi dx \, , \quad \text{giving} \quad E = E_0 \, \exp\left(\frac{-x}{1/N\Phi}\right) \, .$$

Radiation length, usually called X_0, is defined as the thickness of material where an electron will reduce its energy by a factor $1/e$ by bremsstrahlung losses. This corresponds to $1/N\Phi$ in the formula shown above. Radiation length is often parametrized in terms of well-known material properties. A formula which is good to 2.5% (except for helium) is

$$X_0 = \frac{716.4 \, \mathrm{g \, cm^{-2}} \, A}{Z(Z+1) \ln(287/\sqrt{Z})} \, .$$

Multiplying by the density for the various materials we obtain the following: air $\cong$ 300 m, plastic scintillators $\cong$ 40 cm, Si $\cong$ 9 cm, Pb = 0.56 cm, Fe = 1.76 cm.

2.3 Photons

Photons are important for many reasons. They appear in detector systems as primary photons; they are created in bremsstrahlung and de-excitations; and they are used for medical applications, both imaging and radiation treatment.

They react in matter by transferring all (or most) of their energy to electrons, which then lose energy as described above. A beam of photons therefore does not lose energy gradually; it is attenuated in intensity (only partly true because of Compton scattering). Three processes dominate the photon

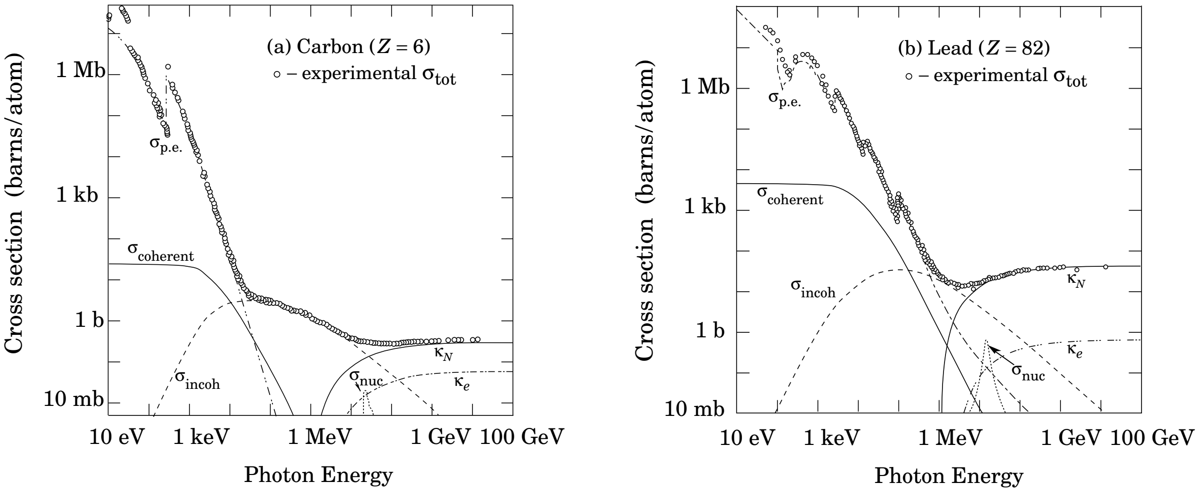


Fig. 4: Dominant processes in photon energy loss

energy loss: 1) Photoelectric effect (goes roughly as Z^5): absorption of a photon by an atom ejecting an electron. The cross-section shows the typical shell structures in an atom. 2) Compton scattering (Z): scattering of an electron against a free electron (Klein NishinaŠs formula). This process has well-defined kinematic constraints (giving the 'Compton Edge' for the maximum energy transfer to the electron) and for energies above a few MeV 90% of the energy is transferred. 3) Pair production ($Z^2 + Z$): essentially the bremsstrahlung process again with the same machinery as used earlier, with a threshold at $2m_e$ = 1.022 MeV. As with bremsstrahlung for electrons, this process dominates at high energies. The most significant processes are shown in Fig. 4 (from Ref. [3]).

Considering only the dominating effect at high energy, the pair-production cross-section, we can calculate the mean free path of a photon based on this process alone:

$$\lambda^{\text{photon}} = \frac{\int x \exp(-N\sigma_{\text{pair}} x)dx}{\int \exp(-N\sigma_{\text{pair}} x)dx} \cong \frac{9}{7} X_0 \ .$$

This shows that around one radiation length is a typical thickness for both bremsstrahlung losses (by 1/e) and pair-production processes.

2.4 Electromagnetic calorimeters

By considering only bremsstrahlung and pair production, dominating at energies above a few tens of MeV, with one splitting per radiation length (either bremsstrahlung or pair production), we can extract a good model for electromagnetic (EM) showers. In such a model the number of tracks increases with the number of radiation lengths t as $N(t) = 2^t$. The energy carried by each particle decreases as $E(t) = E_0/2^t$. This process stops as the energy reduces to the critical energy E_C. After this point the dominating processes are ionization losses, Compton scattering, and photon absorption. From this, the following simple relations can be extracted: the maximum number of tracks, i.e., the shower maximum, is reached at $t_\text{max} = \ln(E_0/E_\text{C})/\ln 2$. The total number of tracks T is $2^{(t_\text{max}+1)} - 1 \approx 2E_0/E_\text{C}$. The total track length is given by $E_0 X_0/E_\text{C}$. The intrinsic relative resolution of a calorimeter is therefore improving with energy:

$$\frac{\sigma(E)}{E} \propto \frac{\sigma(T)}{T} \propto \frac{1}{\sqrt{T}} \propto \frac{1}{\sqrt{E}} \ .$$

Furthermore, the depth needed to contain the shower increases only logarithmically.

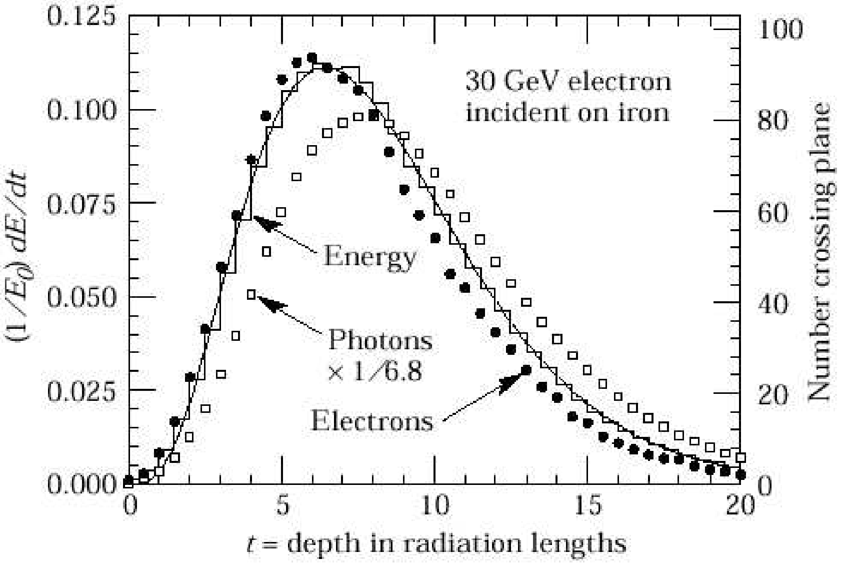


Fig. 5: Energy loss profile, measured and simulated, of electrons and photons (from Ref. [3])

In reality calorimeter resolutions are parametrized also with additional terms to take into account effects of inhomogeneities, cell intercalibrations, non-linearity, and electronics noise and pile-up (a constant term and a $1/E$ term).

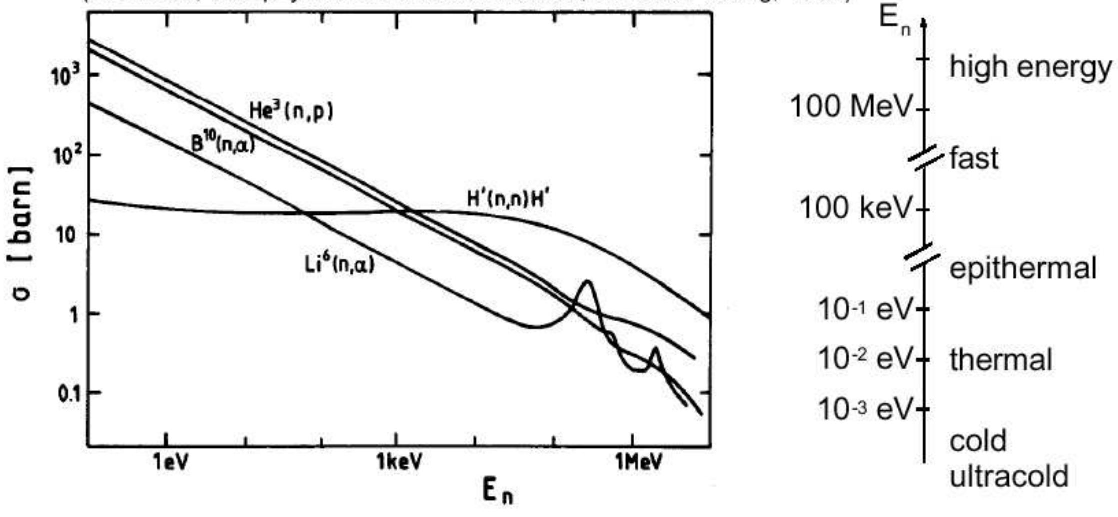
The typical EM shower is 95% contained in a transverse cylinder with radius $2R_{\text{m}}$ = 21 MeV X_0/E_{C}; which should be compared to full longitudinal containment which requires around $25X_0$.

The best performance of EM calorimeters is traditionally achieved with homogeneous crystal calorimeters; typical examples are BGO, CsI, NaI and PWO. The radiation lengths of these materials are 1–2 cm. Drawbacks are cost, radiation effects, and temperature dependence. Sampling calorimeters are often used in large calorimeter systems, where a fraction of the total energy is sampled and the functions of particle absorption (often Pb) and shower sampling (scintillators, ionization detectors, silicon) are separated.

2.5 Neutrons, hadronic absorption/interaction length and hadronic showers

Neutrons have no charge and interact with matter through the strong nuclear force. They transfer energy to charged particles by elastic scattering against protons (below 1 GeV), and are absorbed/captured in materials below 20 MeV (see Fig. 6). Above 1 GeV hadronic cascades are created. We can define hadronic absorption and interaction lengths by the mean free path of hadrons, using the inelastic or total cross-section for high-energy hadrons (above 1 GeV the cross-sections vary little for different hadrons or energy). This is in analogy to the relation between the radiation length and the mean free path of a high-energy photon. In Table 1 (extracted from Ref. [3]) radiation lengths and interaction lengths for various materials are listed.

Hadronic calorimeters usually have a thickness of around 7–8 hadronic interaction lengths (Fig. 7). Their resolution is worse than that of electromagnetic calorimeters for a variety of reasons: there are significant fluctuations between the electromagnetic ($\pi_0 \to 2\gamma$) and hadronic parts (mostly charged pions) of the showers which have to be dealt with, a significant amount of the hadronic energy is lost in break-up of nuclear bindings, muons and neutrinos are created in the shower escaping partly or fully, etc. The key element for good hadronic calorimeters is therefore to understand and minimize the differences between neutral-pion (i.e. photons) and charged-pion response. Several methods are used: compensation, use of tracking information, and use of longitudinal sampling information. Good coverage, uniform



(H. Neuert, Kernphysikalische Messverfahren, G. Braun Verlag, 1966)

Fig. 6: Cross-sections for various neutron processes. The reference is shown in figure.

Table 1: Radiation and interaction lengths for various materials

Material	Z	A	ρ (g/cm^3)	X_0 (g/cm^2)	Λ (g/cm^2)
Hydrogen (gas)	1	1.01	0.0899 (g/l)	6.3	50.8
Beryllium	4	9.01	1.848	65.2	75.2
Silicon	14	28.09	2.33	22	106.4
Iron	26	55.85	7.87	13.9	131.9
Lead	82	207.19	11.35	6.4	194.0

response and adequate granularity in depth and in angular coverage are other important parameters for hadronic calorimetry.

2.6 Neutrinos

Neutrinos react very weakly with matter. For example, the cross-section for $\nu_e + n \rightarrow e^- + p$ above a few MeV is around 10^{-43} cm^{-2} which means that in 1 m of iron the reaction probability is 10^{-17}. Neutrino experiments are therefore very massive and require high fluxes.

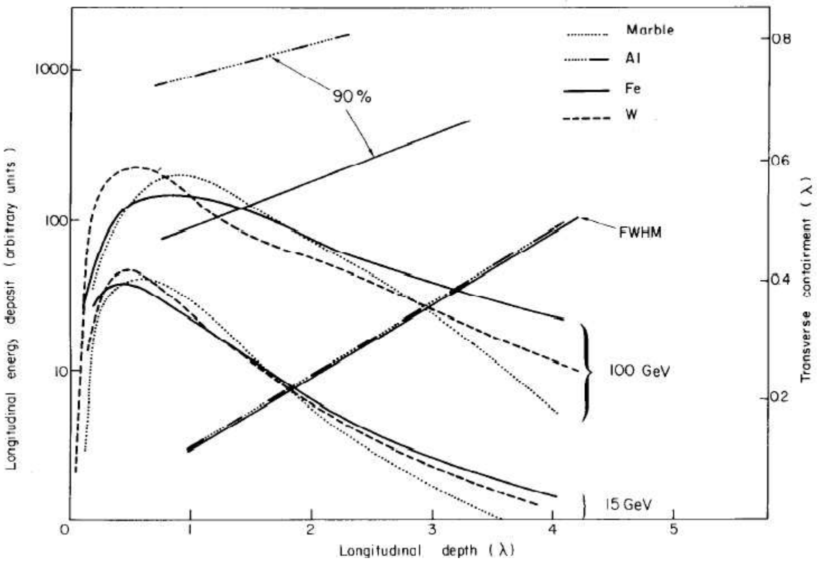
In collider experiments, fully hermetic detectors allow neutrinos to be detected indirectly. The recipe is as follows:

- Sum up all visible energy and momentum in the detector.
- Attribute missing energy and momentum to the escaping neutrino.

The most typical example is the UA1 and UA2 discoveries of $W \rightarrow e\nu$ where this method was used.

3 Particle identification, magnetic fields and combined detector configurations

Section 2 summarized how most 'stable' particles react with matter. We are interested in all important parameters of the particles produced in an experiment: momentum, energy, velocity, charge, lifetime and particle type.



(C. Fabjan, T. Ludlam, CERN-EP/82-37)

Fig. 7: Hadronic shower profiles for hadrons in various materials. The reference is shown in figure.

In the current section we shall look at some specific measurements where 'special effects' or optimized detector configurations are used. Cherenkov and transition radiation are important in detector systems since these effects can be used for particle identification and tracking, even though the energy loss is small. This naturally leads to particle identification with various methods: dE/dx, Cherenkov, transition radiation tracker, electromagnetic and hadronic (EM/HAD), p/E. Secondary vertices/lifetime measurements and combinatorial analysis provide information about c, b-quark systems, taus, converted photons, neutrinos, etc. Finally we shall look at magnetic systems and multiple scattering.

3.1 Cherenkov radiation

A particle with velocity $\beta = v/c$ in a medium with refractive index n may emit light along a conical wave front if the speed is greater than the speed of light in this medium: c/n. The angle of emission (see Fig. 8) is given by

$$\cos\theta = \frac{c/nt}{\beta ct} = \frac{1}{\beta n}$$

and the number of photons by

$$N[\lambda_1 \to \lambda_2] = 4.6 \cdot 10^6 \left[\frac{1}{\lambda_2(A)} - \frac{1}{\lambda_1(A)}\right] L(\mathrm{cm})\, \sin^2\theta\,.$$

In many cases, a Cherenkov threshold detector is used to identify particles of a special type, typically electrons in a beamline. The Cherenkov angle will vary from slightly above 1 degree in the case of air to above 45 degrees for quartz. Generally, by measuring this angle the speed of the particle can be measured. When combined with momentum information, this provides a powerful particle identification tool. The number of photons is small and furthermore one has to take into account detection efficiency of the photons. The goal is to reconstruct a ring in order to provide a measurement of the emission angle and hence the β of the particle.

An example of the DELPHI ring imaging Cherenkov system is shown in Fig. 9. This is a very sophisticated detector which combines a liquid (C_6F_{14}) and a gas radiator (C_5F_{12}/C_4F_{10}), together with a photon detector (TMAE).

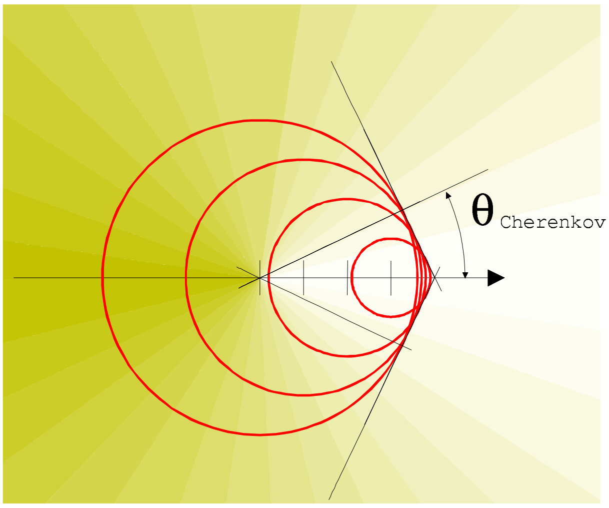


Fig. 8: Cherenkov radiation

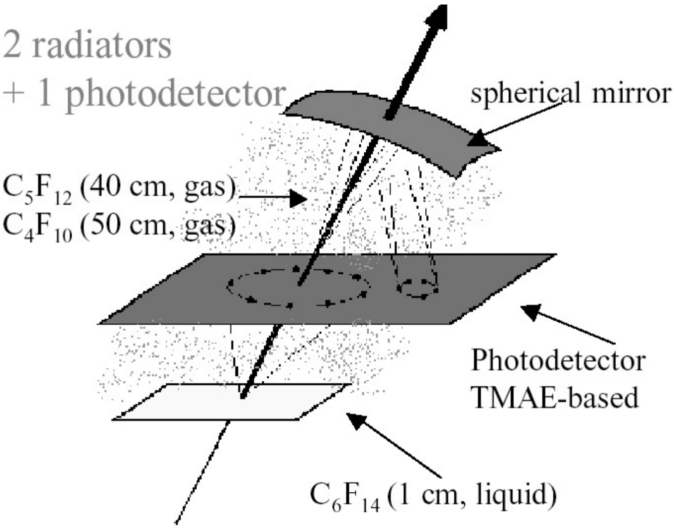


Fig. 9: Principle of ring imaging Cherenkov detector (DELPHI) from Ref. [4], showing the geometrical set-up that allows measurement of the Cherenkov angle. The photon detector must have a high efficiency and be built out of light materials, and hence it is a significant challenge in itself.

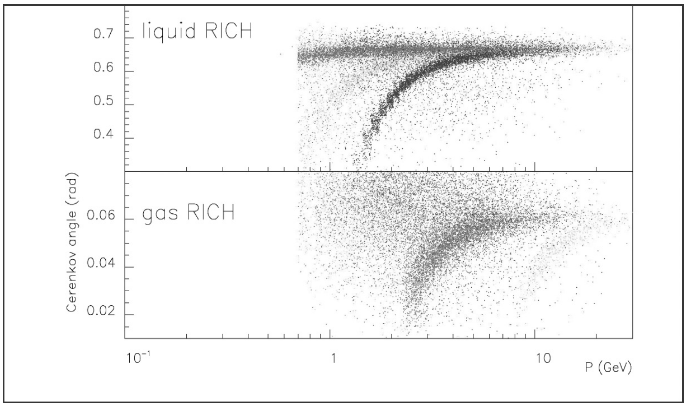


Fig. 10: Cherenkov angle in radians as a function of momentum (GeV) for the DELPHI ring imaging Cherenkov detector. The data in blue are p from Λ, in green K from ΦD^*, in red p from K_0.

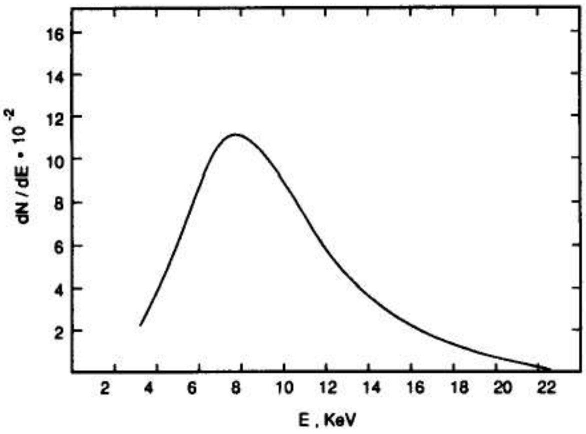


Fig. 11: Simulated spectrum from transition radiation in a stack of CH_2 foils (from Ref. [2]).

3.2 Transition radiation

Electromagnetic radiation is emitted when a charged particle traverses a medium with discontinuous refractive index, as the boundary between vacuum and a dielectric layer. More details can be found in Ref. [5].

The number of photons is small, so many transitions are needed. Hence a stack of radiation layers is interleaved with active detector parts. The emission is proportional to γ, so only high-energy electrons and positrons will emit transition radiation. The energy per boundary is given by

$$W = \frac{1}{3}\,\alpha\hbar\omega_p\gamma$$

and the plasma frequency for a plastic radiator:

$$\hbar\omega_p = \hbar\sqrt{\frac{N_e e^2}{\varepsilon_0 m_e}} \approx 20 \text{ eV}\,.$$

The keV range photons ($\frac{1}{4}\,\hbar\omega_p\gamma$, see Fig. 11) are emitted at a small angle: $\theta \propto 1/\gamma$. The number of photons can be estimated as $W/\hbar\omega_p\gamma \propto \alpha$. The radiation stack has to be transparent to these photons (low Z), hence hydrocarbon foams and fibre materials are used. The detectors have to be sensitive to the photons [high Z, for example Xe (Z = 54)] and at the same time be able to measure dE/dx of the 'normal' particles which have significantly lower energy deposition.

3.3 Particle identification with *dE/dx*

Going back to the Bethe–Bloch plot in Fig. 1, one can see that particles with different masses will in a certain momentum range have different average energy-loss. This is exploited to identify particles. The dE/dx measurements are used to identify particles at relatively low momentum. Figure 12 shows data from the PEP4 time projection chamber with 185 samples (many samples required to handle statistical fluctuations). It can be seen that this method provides efficient particle identification in this momentum range.

3.4 Momentum measurements in a magnetic field and multiple scattering

Consider a particle with charge q and transverse momentum p_{T} moving in a uniform magnetic field B going into the transverse plane, over a length L. The relation between the transverse momentum p_{T} and the radius of curvature ρ is given by $p_{\mathrm{T}} = qB\rho$. Expressing momentum in GeV/c and the magnetic field in tesla (T), and considering q equal to the elementary charge, this gives p_{T} (GeV/c) = 0.3 $B\rho$ (T m).

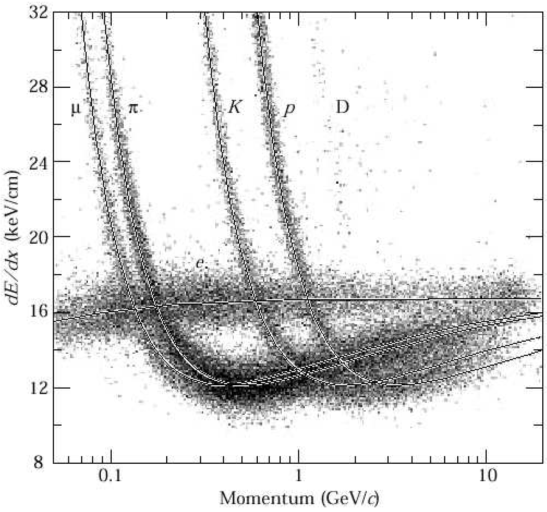


Fig. 12: The dE/dx measured in the PEP4 time projection chamber (Ref. [3]).

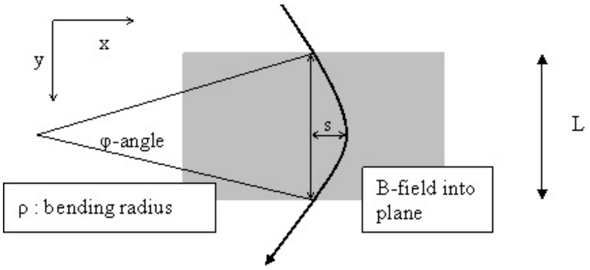


Fig. 13: Bending of a charged particle in magnetic field B

3.4.1 *Measuring the momentum*

Since ρ is much larger than L, as can be seen from the formula above for particles in the GeV range, we can (see Fig. 13) extract the following relations between the sagitta s and the transverse momentum p_T:

$$\begin{aligned} \frac{L}{2\rho} &= \sin\frac{\theta}{2} \approx \frac{\theta}{2}\,, \\ s &= \rho\left(1-\cos\frac{\theta}{2}\right) \approx \rho\frac{\theta^2}{8} \approx \frac{0.3}{8}\,\frac{L^2 B}{p_\mathrm{T}}\,. \end{aligned}$$

By measuring the sagitta $s = x_2 - (x_1 + x_3)/2$, where x is measured at the entrance, middle, and exit of the field region in Fig. 13, we can therefore measure the p_T of the particle. Furthermore, the measurement precision is given by

$$\frac{\sigma(p_\mathrm{T})}{p_\mathrm{T}} = \frac{\sigma(s)}{s}\,\frac{\sqrt{\frac{3}{2}}\,\sigma(x)}{s} = \frac{\sqrt{\frac{3}{2}}\,\sigma(x)\,8\,p_\mathrm{T}}{0.3BL^2}\,.$$

The measurement uncertainty increases linearly with p_T. If N equidistant measurements are used, the expression becomes (Ref. [6])

$$\frac{\sigma(p_\mathrm{T})}{p_\mathrm{T}} = \frac{\sigma(x)\,p_\mathrm{T}}{0.3BL^2}\sqrt{720/(N+4)}\,, \qquad \text{for} \qquad N \geq 10\,.$$

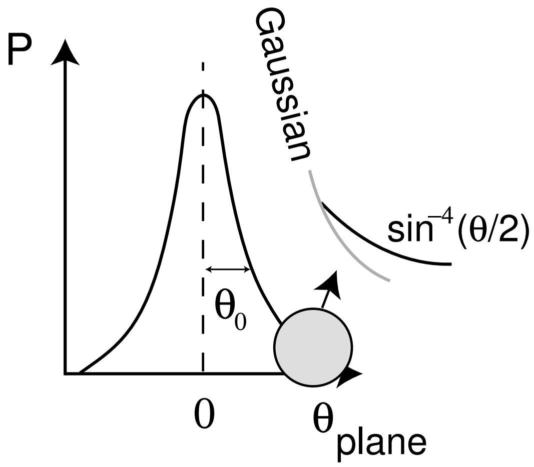


Fig. 14: Gaussian approximation of a multiple scattering distribution, indicating also that the initial Rutherford formula will increase the tails. From Ref. [4].

3.4.2 Multiple scattering processes

These processes will influence the measurement. The cross-section for the scattering between an incoming particle with charge z and a target of nuclear charge Z is given by Rutherford's formula:

$$\frac{d\sigma}{d\Omega} = 4\,z\,Z\,r_e^2\left(\frac{m_e c}{\beta p}\right)^2 \frac{1}{\sin^4\theta/2}\,.$$

For sufficiently thick materials the particle will undergo multiple scattering; usually a Gaussian approximation (Ref. [3]) for the scattering angle distribution is used (see Fig. 14) with a width expressed in terms of radiation length (good to 11% or better):

$$\theta_0 = \frac{13.6\ \text{MeV}}{\beta\,c\,p}\,z\,\sqrt{x/X_0}\,\left[1 + 0.038\,\ln(x/X_0)\right]\,.$$

The multiple scattering over the distance L mentioned above will influence the momentum as follows:

$$\Delta p^{\text{MS}} = p\,\sin\theta_0 \approx 0.0136\sqrt{\frac{L}{X_0}}\,.$$

This should be compared to the change in momentum over the same distance L due to the effect of the magnetic field, see Fig. 13: 0.3 BL.

$$\left.\frac{\sigma(p_{\text{T}})}{p_{\text{T}}}\right|^{\text{MS}} = \frac{\Delta p^{\text{MS}}}{0.3\,BL} = \frac{0.0136\sqrt{\frac{L}{X_0}}}{0.3\,BL} = 0.045\,\frac{1}{B\sqrt{L\,X_0}} \qquad \text{independent of } p\,.$$

The resulting total momentum resolution, adding the two contributions in quadrature, is shown in Fig. 15 (from Ref. [4]).

3.5 Vertexing and secondary vertices

Several important measurements in particle physics depend on the ability to tag and reconstruct particles coming from secondary vertices hundreds of microns from the primary (giving track impact parameters in the tens of micron range), in order to identify systems containing b, c, τ, etc., i.e., generally systems with these types of decay lengths.

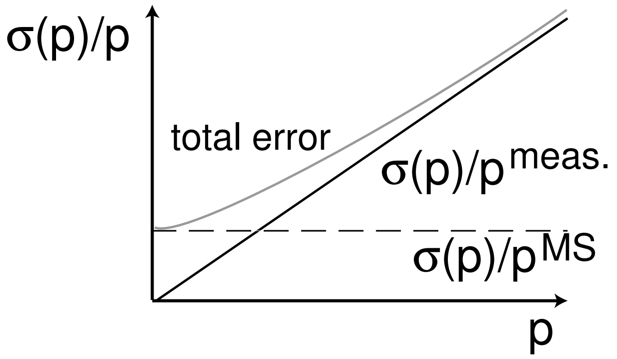


Fig. 15: Total momentum resolution: The measurement uncertainty introduces a linear term in the momentum resolution while multiple scattering introduces a constant term.

This is naturally done with precise vertex detectors where three features are important:

- robust tracking close to vertex area;
- innermost layer as close as possible to the collision point;
- minimum material before first measurement in particular to minimize the multiple scattering (beam pipe most critical).

The vertex resolution is usually parametrized with a term taking into account the geometrical layout of the detector and a term depending on multiple scattering effects, the latter decreasing in importance as the momentum is increased.

3.6 Particle identification combining information from a detector system

In addition to the methods mentioned above, we must keep in mind that combining information from various parts of the detector provides powerful particle identification.

EM/HAD energy deposition information provides particle ID; EM response without a track indicates a photon; matching of p (momentum) and EM energy the same (electron ID); isolation cuts help to identify leptons; vertexing helps us to tag b, c or τ; missing transverse energy indicates a neutrino; muon chamber hits indicate a muon; etc. So, ultimately, a number of combinatorial methods are used in experiments.

4 Active detector elements in particle physics

In Sections 2 and 3 we described how most particles —i.e., all particles that live long enough to reach the detector (electrons, muons, protons, pions, kaons, neutrons, photons, neutrinos, etc.)—react with matter and how they are measured (p, E, v, lifetimes, charge, etc.) and identified in a modern detector system. One essential step in the process was omitted: How are reactions of the various particles with detector elements turned into electrical signals? We want position and energy deposition information, channel by channel, from our detector system.

Three detector types are usually used: ionization detectors, scintillation detectors, and semiconductor detectors. These active elements are used for tracking, energy measurements, or in photon detectors for Cherenkov or TRT. The three types have different applications, advantages and disadvantages, but virtually all active elements in a complex detector system rely on these three principles.

At the end of Section 4 we shall have a quick look at how electrical signals are amplified in front-end electronics, and at the main parameters determining the performance of readout electronics.

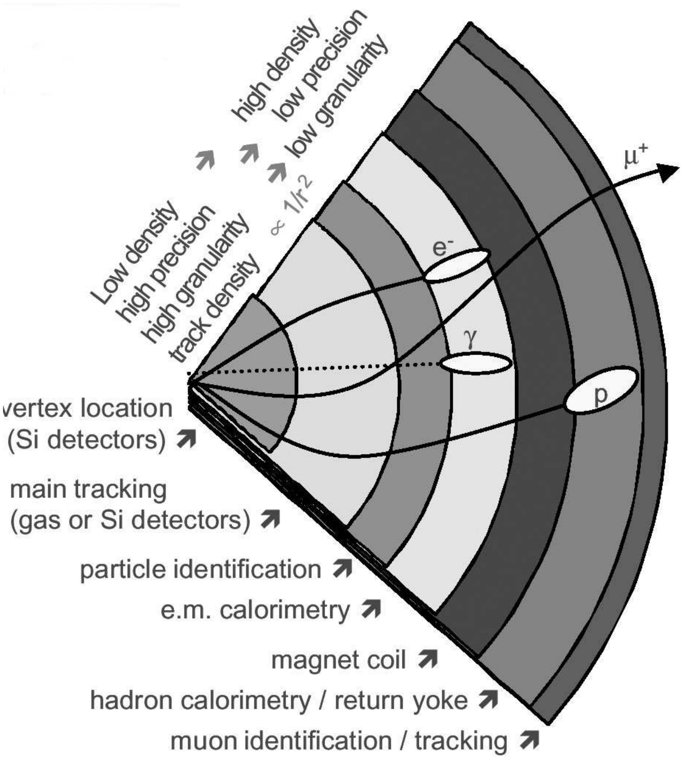


Fig. 16: A typical detector cross-section showing a tracker, particle identification, calorimeters and a muon system, together with the magnets (from Ref. [4])

4.1 Ionization detectors

A charged particle passing through matter will transfer energy to the atomic electrons causing ionization and excitation. In an ionization detector, the electrons and ions created when the particle traverses or is absorbed in a medium, usually gas, are used to generate a measurable signal. The ionization potential for various gases is shown in Fig. 17. Typical numbers of primary encounters in various gases are summarized in Table 2. Since many of these encounters lead to secondary and tertiary ionizations, the number of free electrons created is larger by a factor 3–4; nevertheless, the signal is very small and an amplification step is needed to increase the noise margins.

In the following section the amplification processes and drift in an electrical field are briefly discussed, as they provide the basis for the operation of a proportional chamber. Ionization detectors are generally operated in proportional mode where an amplification of 10^4 to 10^6 is used. The response of a proportional chamber (Fig. 21) is shown in Fig. 18 as a function of voltage. There are several distinctive regions of the response curve:

1. *Recombination* before charge collection.
2. *Ionization* chamber region: all primary charge is collected (no multiplication), giving a flat response.
3. *Proportional* counter (gain up to 10^6), where the electric field is large enough to begin multiplication; secondary avalanches need to be quenched. At the end of this region limited proportionality is observed (secondary avalanches distort the field, more quenching is needed) and the same signal is detected independently of the original ionizing event.
4. *Geiger–Müller* mode, where strong photon emission propagates avalanches all over the wire.

The amplification process can be characterized as follows.
Let α^{-1} be the mean free path (also called the first Townsend coefficient) between each ionization. The

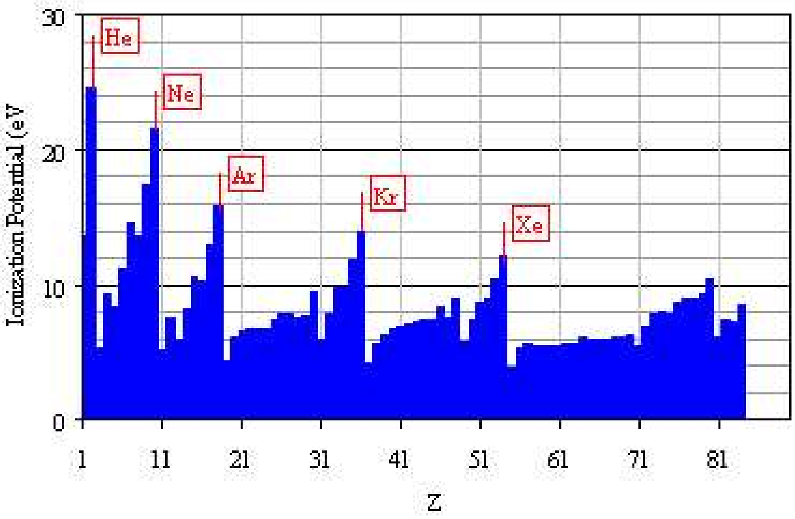


Fig. 17: Ionization potential for various elements

Table 2: Primary and total number of ions/electrons created per centimetre in several gases (at SPT) used in proportional counters

Gas	Primary electrons	Total electrons
	(1/cm)	(1/cm)
He	5	16
Ne	12	42
Ar	25	103
Xe	46	340
CH_4	27	62
CO_2	35	107
C_2H_6	43	113
DME	55	160
i-C_4H_{10}	84	195

increase in the number of produced electrons after a path dx will be $dn = n\alpha dx$, where n is the number of initial electrons. By integration, $n = n_0 e^{\alpha x}$, therefore the gas amplification $M = n/n_0$ is given by

$$M = e^{\int_{x_1}^{x} \alpha(x)dx} .$$

The amplification curve in a standard gas mixture such as Ar–CO_2 [80%–20%] is shown in Fig. 19.

Another important aspect of the ionization chamber is the drift velocity. In a simple formulation, the drift velocity v_D in an electric field E can be written as:

$$v_\mathrm{D} = \frac{e}{m} E \tau ,$$

where τ is the mean free time between collisions (in general a function of the electric field), m the mass and e the charge. Figure 20 shows the drift of electrons under the action of the electric field (superimposed on the normal thermal movements of the gas molecules).

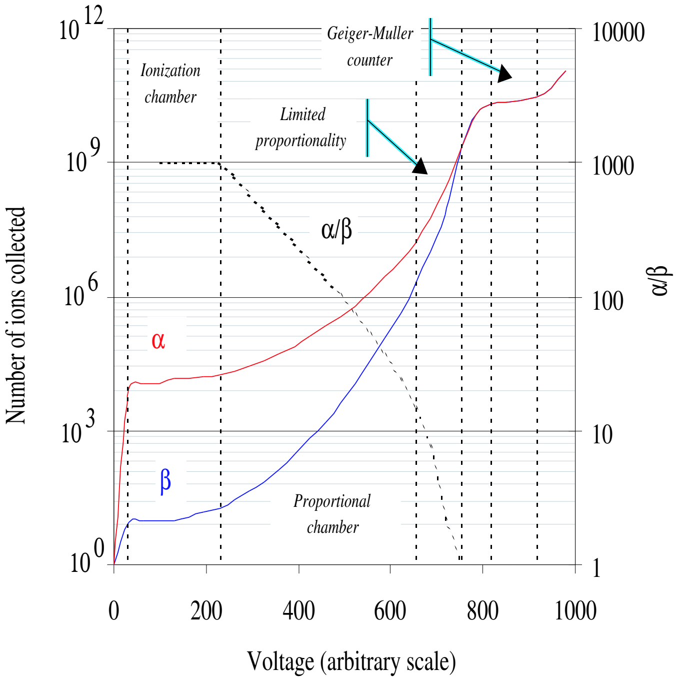


Fig. 18: Response of a proportional chamber as a function of applied voltage (from Ref. [4])

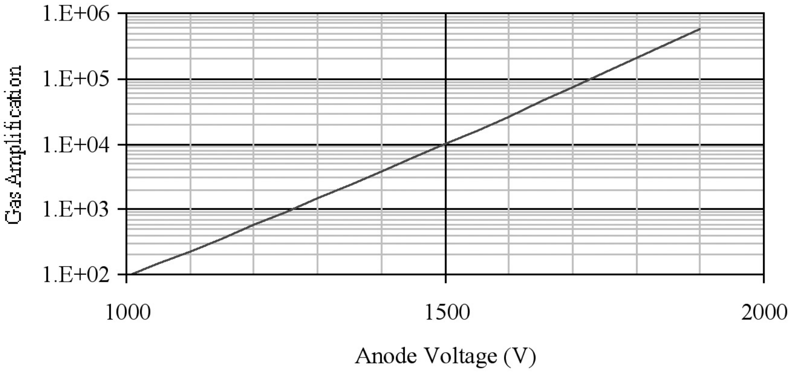


Fig. 19: Gas amplification as a function of voltage in Ar–CO_2 [80%–20%]

Different requirements apply to different chambers. If the chamber is to operate at high counting rates, the drift velocity should be high to avoid losses due to dead time. For better spatial resolution, drift velocities should be lower to minimize the influence of timing errors on position resolution.

In the presence of a magnetic field, the drift velocity is generally reduced, and the drift direction is no longer along the electric field. This has to be taken into account when operating chambers close to or inside strong magnetic fields. The general operational principle of a gas detector can be understood by studying more closely a simple proportional chamber. The cross-section of such a chamber of cylindrical geometry is shown in Fig. 21.

The cathode is a metallic cylinder of radius b. Let us consider a typical example where the anode is a gold-plated tungsten wire of radius a; $a = 10^{-5}$ m and $b/a = 1000$.

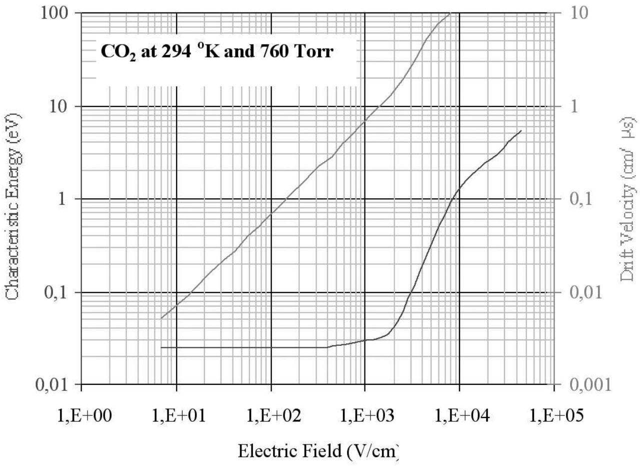


Fig. 20: Drift velocity, upper curve, as a function of electric field for electrons. The drift velocity of the positive ions under the action of the electric field is linear with the reduced electric field (E/pressure) up to very high fields and several orders of magnitude lower than the electron velocity.

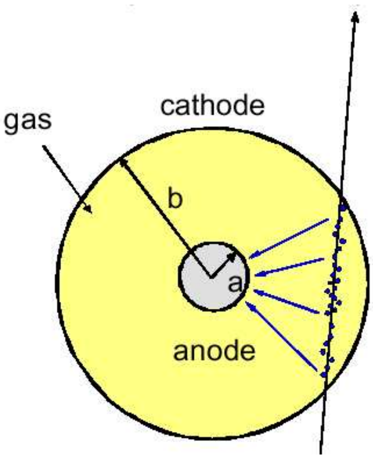


Fig. 21: Cross-section of a proportional chamber

The electric field at a distance r from the centre can be written as

$$E(r) = \frac{1}{r}\,\frac{CV_0}{2\pi\varepsilon_0}\,,$$

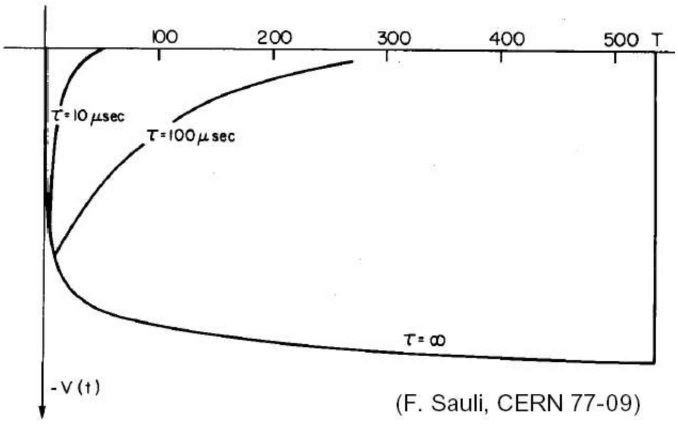
where C is the capacitance per unit length. Given the $1/r$ dependence, the electric field close to the anode is large and multiplication can start; therefore the development of the signal begins at a few wire radii.

The formation of signal can be understood as follows. The electrostatic energy of the configuration is $W = \frac{1}{2}lCV_0^2$, where C is the capacitance per unit length, V_0 the overall potential difference, and l the length of the counter. The potential energy of a charged particle at radius r is given by the charge times the potential:

$$W = -q\,\frac{CV_0}{2\pi\varepsilon_0}\,\ln\frac{r}{a}\,.$$

Considering this as an isolated system, we can set up an equation for how the voltage (signal) changes when the particle moves in the electric field:

$$dW = l\,C\,V_0\,dV = q\frac{d\varphi(r)}{dr}\,dr\,, \qquad \text{where} \quad \varphi(r) = -\frac{C\,V_0}{2\pi\varepsilon}\,\ln\frac{r}{a}\,.$$



(F. Sauli, CERN 77-09)

Fig. 22: Typical signal induced in a proportional chamber. T is the total drift time of positive ions from anode to cathode. The pulse shape obtained with several differentiation time constants is also shown. Electronics differentiation is used to limit dead time. Note that one can speed up the response but at the cost of collecting only a very limited part of the signal. The initial drift time and ultimate time response can be understood from the electron drift in the electrical field and gas mixture used.

The signal is induced mainly by the positive ions created near the anode. This can be seen if we assume that all charges Q are created within a distance λ from the anode. λ is of the order of a few tens of micrometres; hence $V_{\text{electron}} \cong V_{\text{ion}}/100$, which can be seen from the equations below setting in the correct values for a and b:

$$\begin{aligned} V_{\text{electron}} &= -\frac{Q}{lCV_0}\int_a^{a+\lambda}\frac{dV}{dr}dr = -\frac{Q}{2\pi\varepsilon l}\ln\frac{a+\lambda}{a}\,, \\ V_{\text{ion}} &= \frac{Q}{lCV_0}\int_{a+\lambda}^{b}\frac{dV}{dr}dr = -\frac{Q}{2\pi\varepsilon l}\ln\frac{b}{a+\lambda}\,. \end{aligned}$$

The time development of the signal can be computed by neglecting the electron contribution and assuming that all ions leave from the wire surface:

$$V(t) = V_{\text{ion}} = \frac{Q}{lCV_0}\int_{r(0)}^{r(t)}\frac{dV}{dr}dr = -\frac{Q}{2\pi\varepsilon l}\ln\frac{r(t)}{a}\,.$$

The final result for $V(t)$ is shown in Fig. 22. A more general method to look at signal formation is discussed in Section 4.3 (using the Shockley–Ramo theorem).

From the basic proportional chamber, we can now study the following.

- ***Multiwire proportional chambers (MWPC).*** An MWPC consists of a set of thin, parallel anode wires between two cathode planes. The cathodes are at negative voltage and the wires are grounded. This creates a homogeneous electric field in most regions, with all field lines leading from the cathode to the anode wires (Figs. 23 and 24). Multiple planes with different angles of inclination for the wires allow reconstruction of trajectories in space.
  Limited both by electrostatic forces and construction technology, the minimum distance in a MWPC is $\sim$ 1 mm, restricting spatial resolution and rate capability. The binary readout resolution (the r.m.s. of a square probability distribution) is given by pitch/$\sqrt{12}$. Therefore, for a conventional MWPC built with wires spaced by 1 mm, spatial resolution is limited to 300 μm. Analog readout and charge sharing, as shown in Fig. 23 with segmented cathode plane readout, can improve this significantly when the left/right signal size provides detailed information about the hit position. In this case the resolution is limited mainly by the charge sharing mechanisms

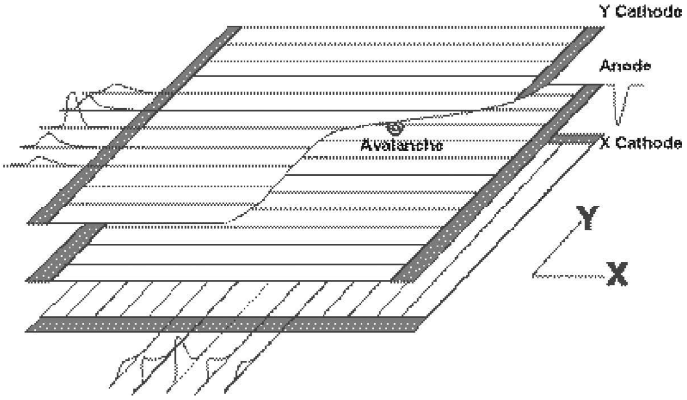


Fig. 23: The basic structure of a two-dimensional MWPC. The avalanche occurring on the anode induces signals of opposite polarity upon the two orthogonal cathode planes. These signals are then used to produce an X and Y position of the incident particle. In general, two-dimensional readout can be obtained by charge division with resistive wires, measurement of timing differences or segmented cathode planes with analog readout as shown here. From Ref. [4].

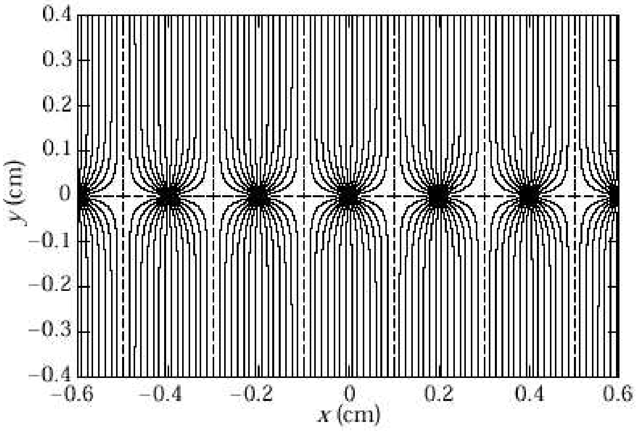


Fig. 24: Electric field equipotentials and field lines in a classic MWPC, from Ref. [4]. Typical parameters: gap between anode and cathode planes, 5 mm; wire spacing, 1–4 mm; anode wire diameter, 20 μm.

and the analog readout resolution. These considerations apply equally well to the silicon detectors discussed in Section 4.3.

- ***Straw tubes.*** The proportional chamber described above, if of small diameter, typically $<$ 10 mm, is a perfect straw-detector unit. Among other advantages, some virtues of a straw system are the possibility of building large self-supporting structures, isolation of broken wires from their neighbours, and minimum cross-talk between neighbouring detector elements.
- ***Drift chambers*** function in the same way as proportional tubes, with measurement of drift time added (time that electrons take to arrive at a sense wire, with respect to a measurement) to determine one coordinate. The space resolution is therefore not limited to cell size, allowing significant reduction of the number of readout channels. The distance between wires is typically 5–10 cm, giving around 1–2 ms drift time. A resolution of 50–100 μm can be achieved, limited by field uniformity and diffusion. There are, however, more problems with occupancy. Drift information is also often used in straw-tube detectors to improve the resolution.
- ***Time projection chambers (TPCs)*** are optimal chambers including all the features above. They permit full three-dimensional track reconstruction, dE/dx and momentum measurements when

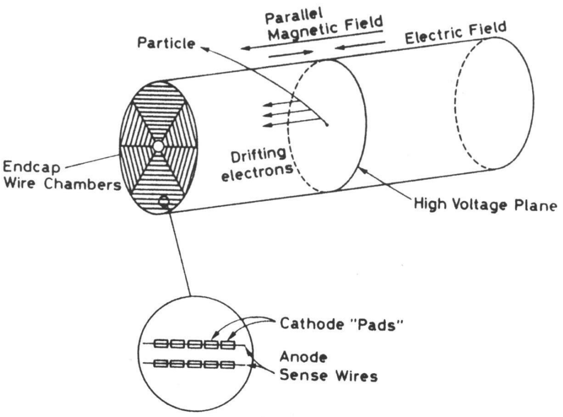


Fig. 25: A typical TPC, centred around the collision point of a collider experiment (from Ref. [2])

used in magnetic fields (see Fig. 25). Their operation is based on the following.

- X and Y coordinates are given by signal readout at the end plate traditionally with conventional MWPC structures.
- Drift-time measurements provide the Z coordinate.
- Analog readout gives dE/dx.
- Magnetic field provides p (and reduces transverse diffusion during drift).

The long drift time and the difficulty of shaping the field are drawbacks: space charge builds up, and inhomogeneities in the field can cause serious degradation of the precision. Introduction of ion-stopping grids (gates), careful tuning of the drift field (sometimes by an additional potential wire plane), and gas purity are of vital importance to the resolution achieved in these chambers.

- Newer chambers and developments such as *Micro Strip Gas Chambers* (MSGC), detectors based on the *Gas Electron Multiplier* (GEM) and the MICRO MEsh GASeous detector (MICROMEGAS) concept.

In recent years there have been several developments directed towards making gas detectors more suitable for high-rate applications, for example as inner-detector components for LHC experiments. MSGCs have been proposed (Ref. [7]) and developed. MSGCs basically reproduce the field structure of MWPCs with a significant scale reduction. They are made of a sequence of alternating thin metallic anode and cathode strips (typical pitch is about 100–200 μm) on an insulating support; a drift electrode on a plane above defines a region of charge collection, and application of appropriate potentials on the strip electrodes creates a proportional gas multiplication field. The intrinsic spatial resolution is about 30 μm r.m.s. using the method of centre of gravity of the amplitude pulses. The multi-track resolution is about 250 μm.

The GEM consists of a thin, metal-clad polymer foil, chemically pierced by a high density of holes (Fig. 26). By applying a potential difference between the two electrodes, electrons released by radiation in the gas on one side of the structure drift into the holes, multiply and transfer to a collection region. The multiplier can be used as a detector on its own, or as a preamplifier in a multiple structure. GEM detectors are used successfully in COMPASS (Ref. [8]). Typical spatial resolution is about 45 μm and time resolution of the order of 12 ns, though lower values can be achieved with suitable gases. Detailed studies of gain and discharge point at high rate, and in the presence of heavily ionizing tracks, have successfully demonstrated the performance of multiple GEM structures in a high-rate environment.

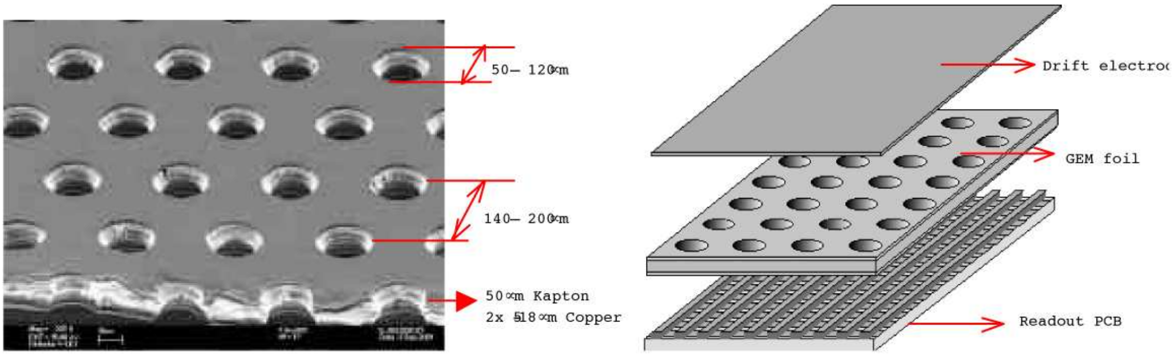


Fig. 26: On the left, SEM picture of a GEM foil. On the right, schematics of a single-GEM detector with two-dimensional readout. The GEM foil separates a drift zone and an induction zone, leading to the readout pad or strip layer. Several GEM foils are used in cascade in some cases.

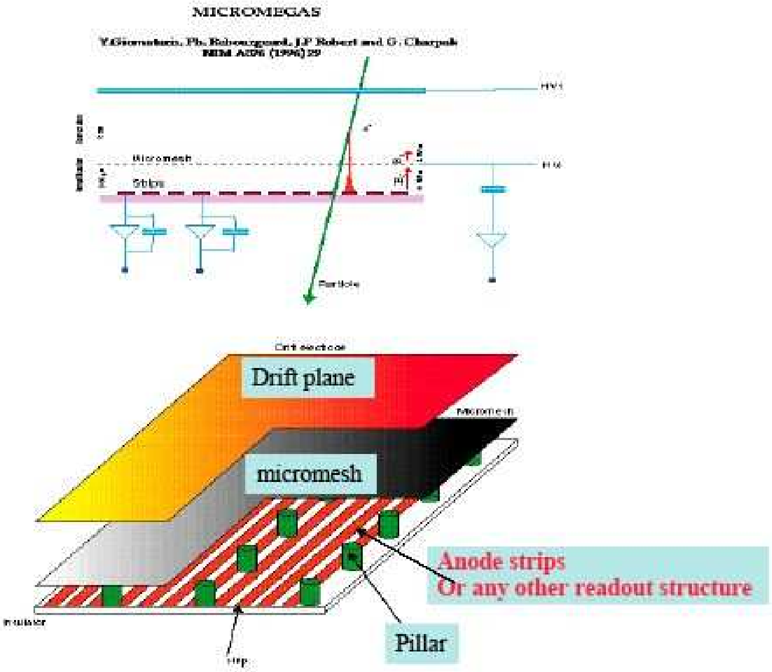


Fig. 27: The MICROMEGAS operation principle

The operational advantages of these developments are based on short drift times and use of PCB and Flex-processing techniques to create the appropriate anode/cathode configurations. A GEM readout for TPCs is also being considered. A GEM-TPC readout end-cap may consist of several cascaded GEMs to obtain the needed amplification, and a patterned readout plane, collecting the (negative) charge.

The MICROMEGAS [9] is a very thin metallic mesh (3–5 μm, Ni or Cu), with a pitch of 20–100 μm located a very small distance from the anode plane (50–100 μm). The very high electric field applied (40–80 kV/cm) creates by avalanche the multiplication of electrons coming from the drift space. These detectors are being used now in several experiments (see, for example, Ref. [10]), and MICROMEGAS readout of TPCs is also being studied. The detectors have excellent two-track separation and spatial resolution, are fast, and operate at high gain. The fabrication technique is also cheap and robust. The concept of the MICROMEGAS is shown in Fig. 27.

For all gaseous detectors the choice of gas is a delicate matter. Gas is selected depending on the desired mode of operation and expected conditions of use. Most chambers run with a mixture of noble gas and a smaller fraction of a polyatomic molecule. The first allows multiplication at low electric fields; the second is chosen because it absorbs photons in a wide energy range, emitted by excited atoms in the avalanche when they return to the ground state, and suppresses secondary emission allowing high gas gains before discharge. A classical gas mixture for low-rate proportional chambers is Ar–CH_4 [90%–10%]. For high-rate, fast detectors, gases with high drift velocity are used to minimize losses due to dead time and occupancy. Better spatial resolution is obtained with low drift velocity gases that minimize timing errors (CO_2 or DME). Microstructures such as MSGCs, GEMs or MICROMEGAS are typically used with gases with high primary ionization statistics to reach full efficiency in thin gas gaps. Finally, radiation damage or ageing of gaseous detectors is a field of continuous study. Experimentally, the progressive loss of detection efficiency or the increase of leakage current in the operating chamber will be interpreted as a clear sign of ageing. These effects depend on many parameters, including gas choice, gas purity and cleanliness, additives and level of impurities, flow rate, gas gain, and detector geometry. Therefore, intensive R&D is needed to set the conditions needed to secure stable operation of gaseous detectors, especially in high-luminosity experiments.

4.2 Scintillators

In scintillating materials, the energy loss of a particle leads to an excitation, quickly followed by a de-excitation providing detectable light. Light detection/readout is therefore an important aspect of the readout of scintillators.

Scintillators are used in many physics applications. They are frequently used in calorimetry (relatively cheap and with good energy resolution), for tracking (fibres), in trigger counters, for time-of-flight measurements, and in veto counters.

Inorganic scintillators are often used in calorimeters because of their high density and Z. They are relatively slow but have high light output and hence good resolution. Organic scintillators are faster but have lower light output. In the following section, both types are discussed further. To convert the light into an electrical signal, a chain of wavelength shifters and photon detectors is used. In this field there are constantly new developments in order to increase granularity, reduce noise and increase sensitivity.

4.2.1 Inorganic crystalline scintillators

The most common inorganic scintillator is sodium iodide (NaI) activated with a trace amount of thallium [NaI(Tl)]. NaI has a light output of typically 40 000 photons per MeV energy loss. The light collection, and the quantum efficiency of the photodetector will reduce the signal further. The detector response is fairly linear. Table 3 lists some commonly used scintillators.

4.2.2 Organic scintillators

These scintillators are fast and with typical light output around half that of NaI. Practical organic scintillators use solvents; typically organic solvents which release a few per cent of the excited molecules as photons (polystyrene in plastic for example, xylene in liquids) + large concentration of primary fluor which transfers to wavelengths where the scintillator is more transparent and changes the time constant + smaller concentration of secondary fluor for further adjustment + … (see Fig. 29).

4.2.3 Light collection and readout

External wavelength shifters and light guides are used to aid light collection in complicated geometries (Fig. 30). These must be insensitive to ionizing radiation and Cherenkov light.

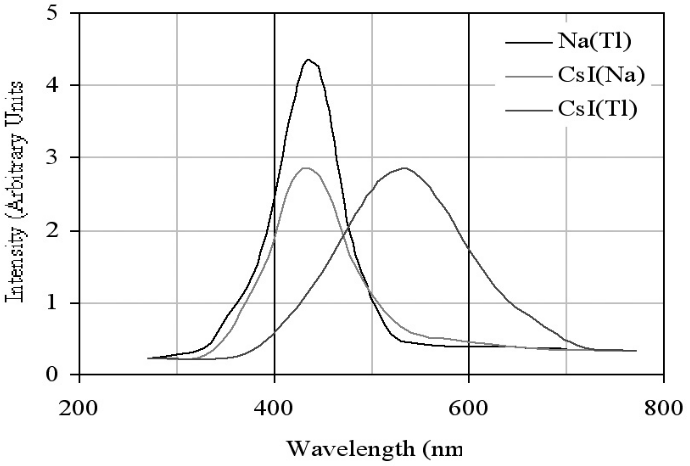


Fig. 28: The light output for NaI and CsI (Ref. [11])

Table 3: A list of the most significant parameters for commonly used scintillators (from Ref. [2])

Crystal	ρ	X_0	$r_{\text{Molière}}$	dE/dx	λ_I	τ_{decay}	λ_{max}	n_{D}	**Rel.**	**Hygro**
	(g/cm^3)	(cm)	(cm)	(MeV/cm)	(cm)	(ns)			**output***	
NaI(Tl)	3.67	2.59	4.5	4.8	41.4	250	410	1.85	1.00	very
BGO	7.13	1.12	2.4	9.2	22.0	300	410	2.20	0.15	no
BaF$_2$	4.89	2.05	3.4	6.6	29.9	0.7	220	1.56	0.05	slightly
						620	310		0.20	
CsI(Tl)	4.53	1.85	3.8	5.6	36.5	1000	565	1.80	0.40	some
CsI(pure)	4.53	1.85	3.8	5.6	36.5	10.36	305	1.80	0.10	some
						36, 620	$\sim$ 480		0.20	
PbWO$_4$	8.28	0.89	2.2	13.0	22.4	5–15	420–440	2.3	0.01	no
CeF$_3$	6.16	1.68	2.6	7.9	25.9	10–30	310–340	1.68	0.10	no

* The light output values are normalized to NaI.

The most critical readout parameters for photodetectors are granularity, noise, and sensitivity. Compared to the typical single-channel photomultipliers (PMs), diodes and triodes, there are several new developments. One example, the multi-anode PM, is shown in Fig. 31. Recently, hybrid photo diodes [12] have been developed where the dynode structure is replaced by a voltage gap and a granular silicon detector (see Fig. 31). This has the potential of removing the primary source of noise, fluctuations in the first dynode, and provides good granularity.

4.3 Solid-state detectors

Solid-state detectors have been used for energy measurements for a long time (silicon, germanium). It takes a few electronvolts to create an electron–hole (e–h) pair and as a result these detector materials have excellent energy resolution. Nowadays silicon detectors are mostly used for tracking and virtually every major particle physics experiment uses this technology for tracking close to the interaction point. We shall concentrate on silicon in the following.

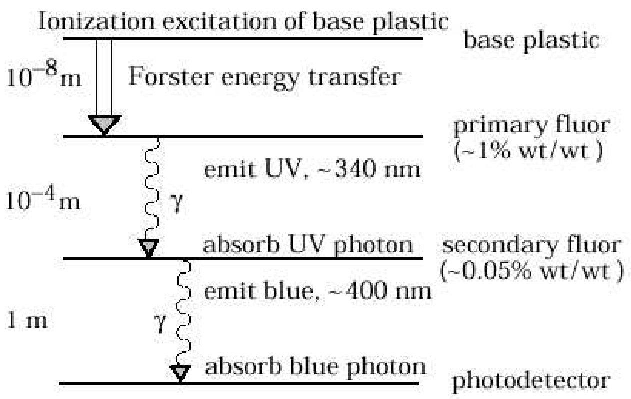


Fig. 29: Illustration of the de-excitation process, including wavelength-shifting fluors, of an organic scintillator (from Ref. [3])

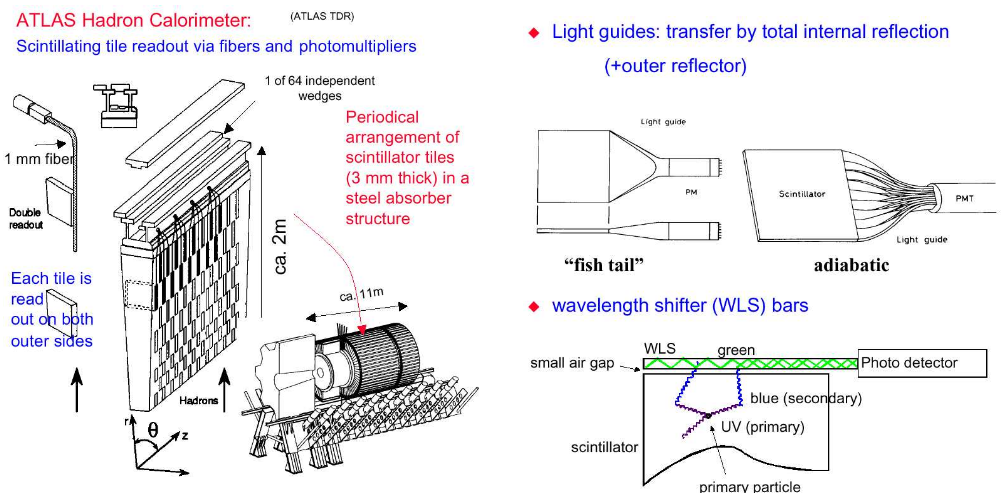
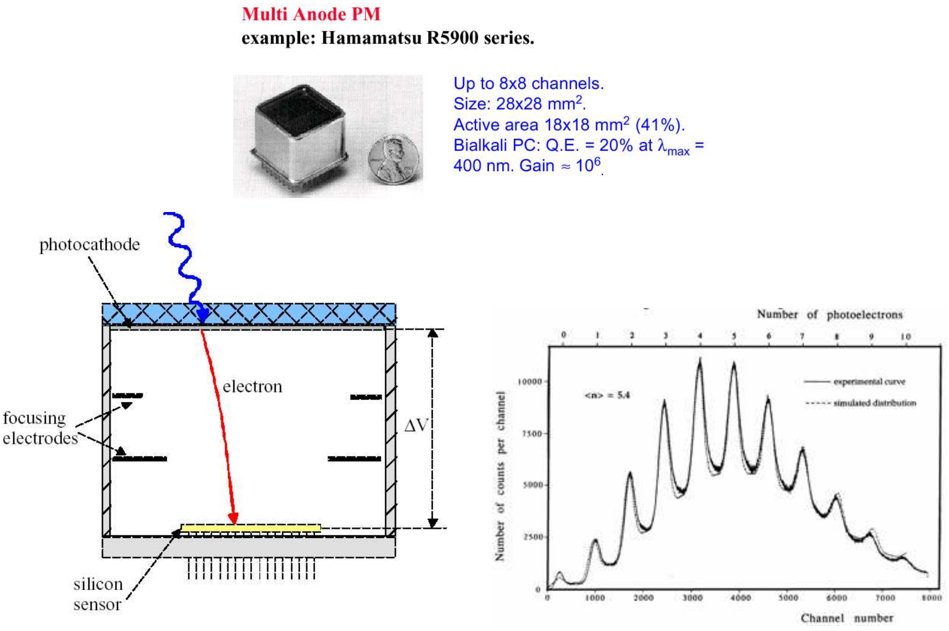

ATLAS Hadron Calorimeter:
(ATLAS TDR)
Scintillating tile readout via fibers and photomultipliers
1 of 64 independent wedges
Periodical arrangement of scintillator tiles (3 mm thick) in a steel absorber structure
1 mm fiber
Double readout
ca. 2m
Each tile is read out on both outer sides
ca. 11m
Hadrons
r
θ
z
Light guides: transfer by total internal reflection
(+outer reflector)
Light guide
PM
Scintillator
PMT
Light guide
"fish tail"
adiabatic
wavelength shifter (WLS) bars
WLS
green
small air gap
Photo detector
blue (secondary)
UV (primary)
scintillator
primary particle

Fig. 30: In the two figures on the right a typical readout configuration is shown, with light guides, wavelength shifter and photodetectors. The ATLAS hadronic calorimeter (left) is a typical example of a modern sampling calorimeter fully based on scintillators as active medium.

The key parameters for silicon detectors are as follows: band gap 1.1 eV, whereas the average energy to create an (e–h) pair is 3.6 eV (compared to 30–40 eV for ionization detectors); high density such that the energy loss in silicon, from Bethe–Bloch, is 108 (e–h)/μm. The mobility for electrons and holes is high, and the structures are self-supporting. More generally, the successful development of modern silicon detectors relies on the progress in the semiconductor industry in recent decades. This concerns key parameters such as reliability, yield, cost, feature sizes, and connectivity.

Contrary to the ionization detectors there is no amplification mechanism, however signal/noise levels of 10–50 are common, mostly depending on the electronics noise, again depending on detector geometry (capacitive load seen by readout amplifier).

Intrinsic silicon will have electron density = hole density: $1.45 \cdot 10^{10}$ cm^{-3} at room temperature (from basic semiconductor theory). In the volume shown in Fig. 32 this would correspond to $4.5 \cdot 10^{8}$ free charge carriers, compared to around $3.2 \cdot 10^{4}$ produced by a minimum-ionizing particle passing it (corresponding to the Bethe–Bloch energy loss in 300 μm Si divided by 3.6 eV). As a result there is a need to decrease the number of free carriers. This is done by using the depletion zone between two oppositely doped parts of a silicon wafer.



Multi Anode PM
example: Hamamatsu R5900 series.

Up to 8x8 channels.
Size: 28x28 mm^2.
Active area 18x18 mm^2 (41%).
Bialkali PC: Q.E. = 20% at λ_{max} = 400 nm. Gain ≈ 10^6.

Fig. 31: Examples of new readout developments for photons aimed at increasing granularity and resolution (from Ref. [3])

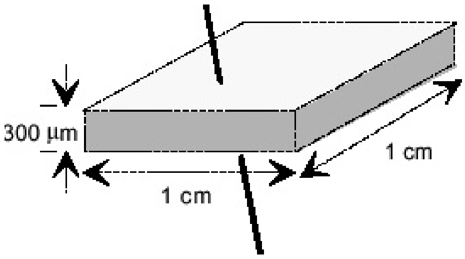


Fig. 32: Sketch of a silicon detector volume

The zone between the N- and P-type doping is free of charge carriers, has an electric field, and is well suited as detector volume. This zone is increased by applying reverse biasing.

One can quickly establish the most critical parameters for a silicon detector by looking at the p,n junction in Fig. 33. We use Poisson's equation:

$$\frac{d^2V}{dx^2} = -\frac{\rho(x)}{\varepsilon},$$

with charge density from $-x_\mathrm{p}$ to 0 and from 0 to x_n defined by $\rho(x) = \pm e\,N_\mathrm{D/A}$. N_D and N_A are the doping concentrations (donor, acceptor): $N_\mathrm{D}x_\mathrm{p} = N_\mathrm{A}x_\mathrm{n}$. The depletion zone is defined as: $d = x_\mathrm{p} + x_\mathrm{n}$.

By integrating once, $E(x)$ can be determined, by integrating twice, the following two important relations are found:

$$\begin{aligned} V &\propto d^2\,, \\ C &= \varepsilon\frac{A}{d} \propto V^{-1/2}\,. \end{aligned}$$

By increasing the voltage the depletion zone is expanded and the capacitance C decreased, giving decreased electronics noise.

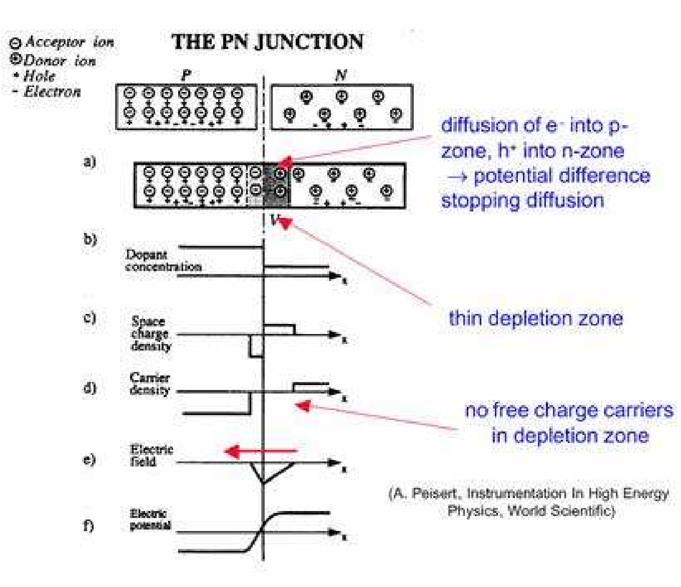

THE PN JUNCTION
⊖ Acceptor ion
⊕ Donor ion
+ Hole
- Electron
P
N
a)
b)
Dopant concentration
c)
Space charge density
d)
Carrier density
e)
Electric field
f)
Electric potential
V
x
diffusion of e⁻ into p-zone, h⁺ into n-zone → potential difference stopping diffusion
thin depletion zone
no free charge carriers in depletion zone
(A. Peisert, Instrumentation In High Energy Physics, World Scientific)

Fig. 33: A silicon p,n junction, see references in figure

Let us have a look at the signal formation using the same simple model of the detector as two parallel electrodes separated by d. Maintaining a constant voltage across the detector with an external bias circuit, an electric charge e moving a distance dx will induce a signal dQ on the readout electrode: $dQ\,d = edx$.

As in the case of the proportional chamber, we use

$$\frac{dx}{dt} = \mu\,E(x)\,,$$

giving (the charge is created at x_0)

$$x(t) = x_0\,\exp\left(\frac{\mu_e t}{\mu_h \tau}\right)\,,$$

where $\tau = \varepsilon/eN_{\mathrm{A}}\,\mu_h$. The time-dependent signal is then

$$Q_e(t) = -\frac{e}{d}\int\frac{dx}{dt}\,dt\,.$$

However, there are many caveats. In reality one has to start from the real (e–h) distribution from a particle. Equally important is to use a real description of $E(x)$ taking into account strips, other implants, and over-depletion, to mention only a few key features. Traps and changes in mobility will also enter.

A more general approach can be used for signal formation. This method applies both to ionization detectors, where we used energy balance to look at how a voltage signal was created due to charge drifting in the device, and to semiconductors as discussed above. More generally we should use the Shockley–Ramo theorem for induced charge:

$$i = q\overrightarrow{v}\cdot\overrightarrow{E}_0$$

or

$$Q = q\Delta\varphi_0$$

where $\overrightarrow{E}_0$ is the weighting field and $\Delta\varphi_0$ the potential difference from the beginning to the end of the path. The weighting potential is found by solving the Laplace equation with some artificial boundary conditions [for the electrode under study (= unity) and for all other electrodes (= 0)] [13].

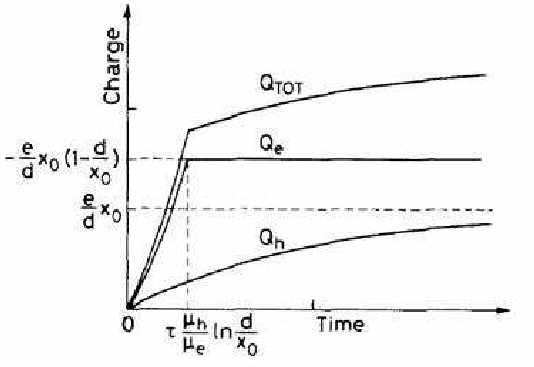


Fig. 34: The final result showing (when entering real numbers and using a more complete model) time-scales of 10–25 ns for electron–hole collection. From Ref. [2].

The main message is that the signal is induced by the motion of charge after incident radiation (not when the charge reaches the electrodes). For ionization chambers it can be used to study not only the signal on the primary anode but also for the neighbours, or the cathode strips (if these are read out). For silicon detectors, it can be used to study charge sharing between strips or pixels.

At the moment silicon detectors are used close to the interaction region in most collider experiments and are exposed to severe radiation conditions (damage).

The damage depends on fluence as well as on particle type (proton, γ, e, neutrons, etc.) and energy spectrum, and influences both sensors and electronics. The effects are due to both bulk damage (lattice changes) and surface effects (trapped charges).

Three main consequences are seen for silicon detectors (figures from Refs. [4, 14]).

- Increase of leakage current with consequences for cooling and electronics. This is illustrated in Fig. 37 on the right.
- Change in depletion voltage, increasing significantly at the end of the detector lifetime; combined with increased leakage currents this leads to cooling problems again (see Fig. 37).
- Decrease of charge collection efficiency.

The future developments for semiconductor systems address four points in particular (Refs. [14–16]).

1. Radiation hardness, cost and power consumption. Examples are the following.
   - Defect engineering: Introduce specific impurities in silicon to influence defect formation.
   - Cooling detectors to cryogenic temperatures.
2. New materials such as diamond and amorphous silicon, the latter opening for deposition directly on readout chips.
3. Integrate the detector and readout on the same wafer.
4. New detector concept as horizontal biasing for faster charge collection and lower biasing voltage. This will also allow the building of detectors which are active very close to the physical edge of the wafers.

4.4 Front-end electronics

A concise description of front-end electronics can be found in Ref. [11]. Here we provide a very short and superficial summary of some of the main concepts and constraints.

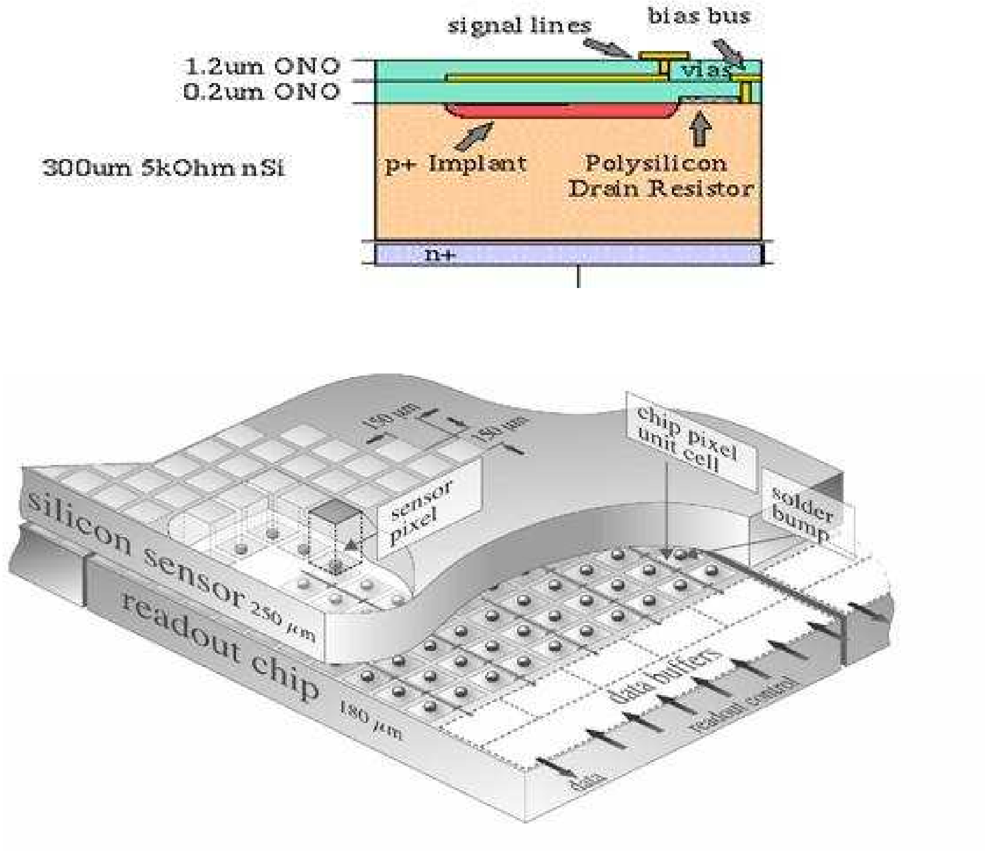


Fig. 35: The detectors used in particle physics are usually strip detectors with strip distance 50–100 μm, single or double sided. One example is shown (top). A more integrated approach is a PIXEL detector (bottom), where the interconnectivity to the readout electronics is made with bump-bonding.

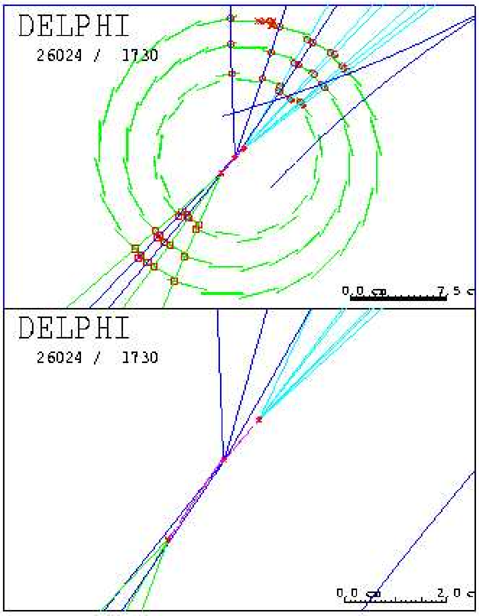


Fig. 36: The microstrip system at LEP was heavily used for B-physics and an example of reconstruction is shown

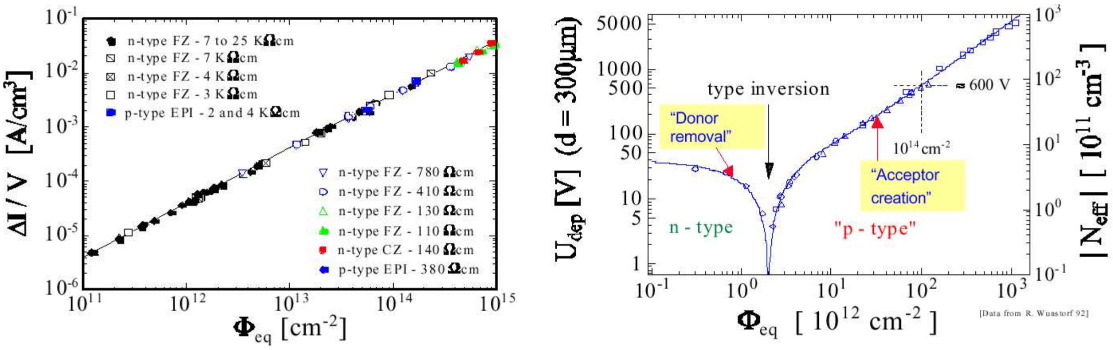


Fig. 37: Change of leakage current and biasing voltage as a function of fluence

Most detectors rely critically on low-noise electronics; optimal detector performance requires optimized electronics solutions.

Figure 38 shows a typical front-end electronics. The detector is represented by the capacitance C_d, bias voltage is applied through R_b, and the signal is coupled to the amplifier through a capacitance C_c. The resistance R_s represents all the resistances in the input path. The preamplifier provides gain and feeds a shaper which takes care of the frequency response and limits the duration of the signal.

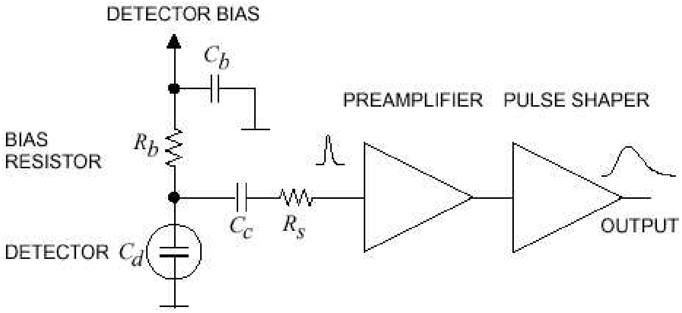


Fig. 38: A typical front-end electronics (Ref. [3])

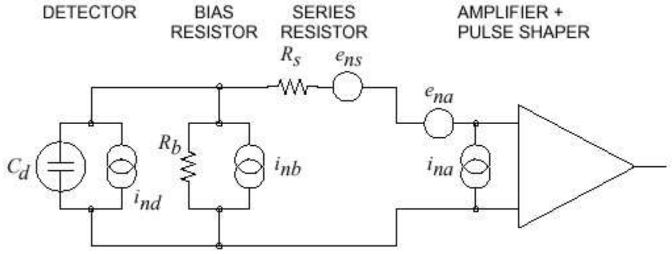


Fig. 39: Front-end electronics for noise analysis. The diagram shows the noise sources and their representation in the noise analysis.

The equivalent circuit for noise analysis (Fig. 39) includes both current and voltage noise sources labelled i_n and e_n, respectively. Two important noise sources are the detector leakage current (fluctuating — sometimes called shot noise) and the electronic noise of the amplifier, both unavoidable and therefore important to control and reduce.

While shot noise and thermal noise have a white frequency spectrum (dP_n/df constant), trapping/detrapping in various components will introduce a $1/f$ noise. Since the detectors usually turn the signal into charge one can express the noise as equivalent noise charge, which is equivalent to the detector signal that yields a signal-to-noise ratio of one.

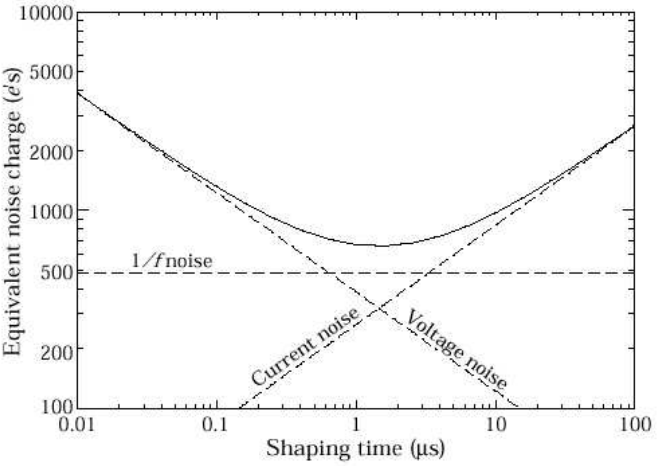


Fig. 40: Optimal shaping time

For the situation we have described there is an optimal shaping time as shown in Fig. 40.

Increasing the detector capacitance will increase the voltage noise and shift the noise minimum to longer shaping times. For quick estimates, one can use the following equation (see Ref. [3]):

$$(Q_\mathrm{n}/e)^2 = 12\left[\frac{1}{\mathrm{nA}\cdot\mathrm{ns}}\right] I_\mathrm{d}\,\tau + 6\times10^5\left[\frac{\mathrm{k}\Omega}{\mathrm{ns}}\right]\frac{\tau}{R_\mathrm{b}} + 3.6\times10^4\left[\frac{\mathrm{ns}}{(pF)^2\,(\mathrm{nV})^2/\mathrm{Hz}}\right] e_\mathrm{n}^2\,\frac{C^2}{\tau}\,,$$

which assumes an FET amplifier with negligible i_na, and a simple CR-RC shaper with time constant τ (equal to the peaking time).

This shows that the critical parameters are detector capacitance, the shaping time τ, the resistances in the input circuit, the leakage current, and the amplifier noise parameters. The latter depends mostly on the input device (transistor) which has to be optimized for the load and use. One additional critical parameter, not apparent in the formula above, is the current drawn which makes an important contribution to the power consumption of the electronics.

Practical noise levels vary between 10^2 and 10^3 ENC for silicon detectors and 10^4 for high-capacitance liquid-argon calorimeters (10^4 corresponds to around 1.6 fC).

5 The future

Improved detectors will certainly be needed. Linear colliders, neutrino facilities, astroparticle physics systems, and LHC upgrades will drive the development and things are already happening. Some main areas of research are the following.

- Radiation hardness will remain a headache: for trackers and calorimeters, active detector elements and electronics, and even for support structures and cooling systems.
- Reduce power or deliver power in a more intelligent way (trackers at the LHC need of the order 100 kW at less than 5 V, currents are huge, cables the same to keep losses acceptable). The services complicate the detector integration and compromise the performance.
- Reduce costs for silicon detectors (strip and various pixels). Today a PIXEL detector costs 5–7 MCHF per m^2, strip trackers around 0.3–1 MCHF per m^2. Similar cost arguments apply to the construction of future large muon chamber systems.
- Learn to work at even higher collision frequency (today 40 MHz for the LHC).

Acknowledgements

A similar set of seminars were given at the CERN European School of High-Energy Physics 2002. For both these sets of seminars the lectures given by C. Joram, CERN, for the 2002 and 2003 CERN Summer Student Lectures [4], and by O. Ullaland, CERN, at the 2001 CERN–CLAF School of High-Energy Physics [11] have provided a lot of detailed information, in particular for Sections 3 and 4. Joram's lectures, being the most comprehensive, contain many practical examples of the use of various detector types and techniques. In several cases comments and pictures from their talks have been used and I can only hope that I have managed to make all the references correctly. Furthermore, since I use Refs. [2] and [3] during my lectures at the University of Oslo for this subject, this summary follows these two sources very closely. Also many thanks to Marko Mikuz for a number of clarifying discussions and corrections to early drafts for the proceedings prepared for the CERN European School of High-Energy Physics in 2002, equally relevant for this summary.

References

[1] K. Kleinknecht, *Detectors for Particle Radiation* (Cambridge Univ. Press, 1998), ISBN-0-521-64032-6; in particular, see Chapter 1.

[2] W.R. Leo, *Techniques for Nuclear and Particle Physics Experiments*, 2nd ed. (Springer-Verlag, Berlin, 1994), ISBN-0-387-57280-5; in particular, see Chapters 2, 6, 7 and 10.

[3] D.E. Groom *et al., Review of Particle Physics;* Section: Experimental Methods and Colliders; see `http://pdg.web.cern.ch/pdg/`
Chapter 27: Passage of particles through matter.
Chapter 28: Particle Detectors.

[4] C. Joram, *Particle Detectors*, CERN Summer Student Lectures 2002 and 2003. These lectures can be found on the Web via CERN's video pages, and they are also video-taped.

[5] B. Dolgosheim, *Nucl. Instrum. Methods Phys. Res.* **A326** (1993) 434.

[6] R.L. Gluckstern, *Nucl. Instrum. Methods* **24** (1963) 381.

[7] F. Sauli, Development of high rate MSGCs: overview of results from RD-28, *Nucl. Phys.* **B61** (1998) 236.

[8] B. Ketzer, Gem detectors for COMPASS, *IEEE Trans. Nucl. Sci.* **NS-48** (2001) 1605.

[9] I. Giomataris *et al., Nucl. Instrum. Methods Phys. Res.* **A376** (1996) 29.

[10] F. Kunne *et al., Nucl. Phys.* **A721** (2003) 1087c-90c.

[11] O. Ullaland, *Instrumentation*, lectures at the 2001 CERN–CLAF School of High-Energy Physics, Itacuruçá, Brazil, CERN Report 2003–003.

[12] C. Joram, *Nucl. Phys. B (Proc. Suppl.)* **78** (1999) 407.

[13] G. Knoll, *Radiation Detection and Measurement*, 3rd ed. (Wiley, New York, 2000), ISBN-0-471-07338-5; in particular Appendix D.

[14] CERN/LHCC 2000-009 LEB Status Report/RD48.

[15] RD39 Collaboration status reports. Available via CERN.

[16] Proceedings Frontier Detectors for Frontier Physics, 9th Pisa meeting on Advanced Detectors, Elba, 25–31 May 2003.

Trigger and data acquisition

N. Ellis
CERN, Geneva, Switzerland

1 Introduction

These lectures concentrate on experiments at high-energy particle colliders, especially the general-purpose experiments at the Large Hadron Collider (LHC). These experiments represent a very challenging case that illustrates well the problems that have to be addressed in state-of-the-art high-energy physics (HEP) trigger and data-acquisition (T/DAQ) systems. This is also the area in which the author is working (on the trigger for the ATLAS experiment at LHC) and so is the example that he knows best. However, the lectures start with a more general discussion, building up to some examples from LEP that had complementary challenges to those of the LHC. The LEP examples are a good reference point to see how HEP T/DAQ systems have evolved in the last few years.

Students at this school come from various backgrounds — phenomenology, experimental data analysis in running experiments, and preparing for future experiments (including working on T/DAQ systems in some cases). These lectures try to strike a balance between making the presentation accessible to all, and going into some details for those already familiar with T/DAQ systems.

1.1 Definition and scope of trigger and data acquisition

The T/DAQ is the online system that selects particle interactions of potential interest for physics analysis (trigger), and that takes care of collecting the corresponding data from the detectors, putting them into a suitable format and recording them on permanent storage (DAQ). Special modes of operation need to be considered, e.g., the need to calibrate different detectors in parallel outside of normal data-taking periods. T/DAQ is often taken to include associated tasks, e.g., run control, monitoring, clock distribution and book-keeping, all of which are essential for efficient collection and subsequent offline analysis of the data.

1.2 Basic trigger requirements

As introduced above, the trigger is responsible for selecting interactions that are of potential interest for physics analysis. These interactions should be selected with high efficiency, the efficiency should be precisely known (since it enters in the calculation of cross-sections), and there should not be biases that affect the physics results. At the same time, a large reduction of rate from unwanted high-rate processes may be needed to match the capabilities of the DAQ system and the offline computing system. High-rate processes that need to be rejected may be instrumental backgrounds or high-rate physics processes that are not relevant for the analyses that one wants to make. The trigger system must also be affordable, which implies limited computing power. As a consequence, algorithms that need to be executed at high rate must be fast. Note that it is not always easy to achieve the above requirements (high efficiency for signal, strong background rejection and fast algorithms) simultaneously.

Trigger systems typically select events[1] according to a ‘trigger menu’, i.e., a list of selection criteria — an event is selected if one or more of the criteria are met. Different criteria may correspond to different signatures for the same physics process — redundant selections lead to high selection efficiency and allow the efficiency of the trigger to be measured from the data. Different criteria may also reflect the wish to concurrently select events for a wide range of physics studies—HEP ‘experiments’ (especially those with large general-purpose ‘detectors’ or, more precisely, detector systems) are really experimental facilities. Note that the menu has to cover the physics channels to be studied, plus additional data

[1] The term ‘event’ will be defined in Section 3 — for now, it may be taken to mean the record of an interaction.

samples required to complete the analysis (e.g., measure backgrounds, and check the detector calibration and alignment).

1.3 Basic data-acquisition requirements

The DAQ system is responsible for the collection of data from detector digitization systems, storing the data pending the trigger decision, and recording data from the selected events in a suitable format. In doing so it must avoid corruption or loss of data, and it must introduce as little dead-time as possible ('dead-time' refers to periods when interesting interactions cannot be selected — see below). The DAQ system must, of course, also be affordable which, for example, places limitations on the amount of data that can be read out from the detectors.

2 Design of a trigger and data-acquisition system

In the following a very simple example is used to illustrate some of the main issues for designing a T/DAQ system. An attempt is made to omit all the detail and concentrate only on the essentials — examples from real experiments will be discussed later.

Before proceeding to the issue of T/DAQ system design, the concept of dead-time, which will be an important element in what follows, is introduced. 'Dead-time' is generally defined as the fraction or percentage of total time where valid interactions could not be recorded for various reasons. For example, there is typically a minimum period between triggers — after each trigger the experiment is dead for a short time.

Dead-time can arise from a number of sources, with a typical total of up to $\mathcal{O}(10\%)$. Sources include readout and trigger dead-time, which are addressed in detail below, operational dead-time (e.g., time to start/stop data-taking runs), T/DAQ downtime (e.g., following a computer failure), and detector downtime (e.g., following a high-voltage trip). Given the huge investment in the accelerators and the detectors for a modern HEP experiment, it is clearly very important to keep dead-time to a minimum.

In the following, the design issues for a T/DAQ system are illustrated using a very simple example. Consider an experiment that makes a time-of-flight measurement using a scintillation-counter telescope, read out with time-to-digital converters (TDCs), as shown in Fig. 1. Each plane of the telescope is viewed by a photomultiplier tube (PMT) and the resulting electronic signal is passed to a 'discriminator' circuit that gives a digital pulse with a sharp leading edge when a charged particle passes through the detector. The leading edge of the pulse appears a fixed time after the particle traverses the counter. (The PMTs and discriminators are not shown in the figure.)

Two of the telescope planes are mounted close together, while the third is located a considerable distance downstream giving a measurable flight time that can be used to determine the particle's velocity. The trigger is formed by requiring a coincidence (logical AND) of the signals from the first two planes, avoiding triggers due to random noise in the photomultipliers — it is very unlikely for there to be noise pulses simultaneously from both PMTs. The time of arrival of the particle at the three telescope planes is measured, relative to the trigger signal, using three channels of a TDC. The pulses going to the TDC from each of the three planes have to be delayed so that the trigger signal, used to start the TDC (analogous to starting a stop-watch), gets there first.

The trigger signal is also sent to the DAQ computer, telling it to initiate the readout. Not shown in Fig.1 is logic that prevents further triggers until the data from the TDC have been read out into the computer — the so-called dead-time logic.

2.1 Traditional approach to trigger and data acquisition

The following discussion starts by presenting a 'traditional' approach to T/DAQ (as might be implemented using, for example, NIM and CAMAC electronics modules[2], plus a DAQ computer). Note that this approach is still widely used in small test set-ups. The limitations of this model are described and ways of improving on it are presented. Of course, a big HEP experiment has an enormous number of sensor channels [up to $\mathcal{O}(10^8)$ at LHC], compared to just three in the example. However, the principles are the same, as will be shown later.

Limitations of the T/DAQ system shown in Fig. 1 are as follows:

1. The trigger decision has to be made very quickly because the TDCs require a 'start' signal that arrives before the signals that are to be digitized (a TDC module is essentially a multichannel digital stop-watch). The situation is similar with traditional analog-to-digital converters (ADCs) that digitize the magnitude of a signal arriving during a 'gate' period, e.g., the electric charge in an analogue pulse — the gate has to start before the pulse arrives.
2. The readout of the TDCs by the computer may be quite slow, which implies a significant dead-time if the trigger rate is high. This limitation becomes much more important in larger systems, where many channels have to be read out for each event. For example, if 1000 channels have to be read out with a readout time of 1 μs per channel (as in CAMAC), the readout time per event is 1 ms which excludes event rates above 1 kHz.

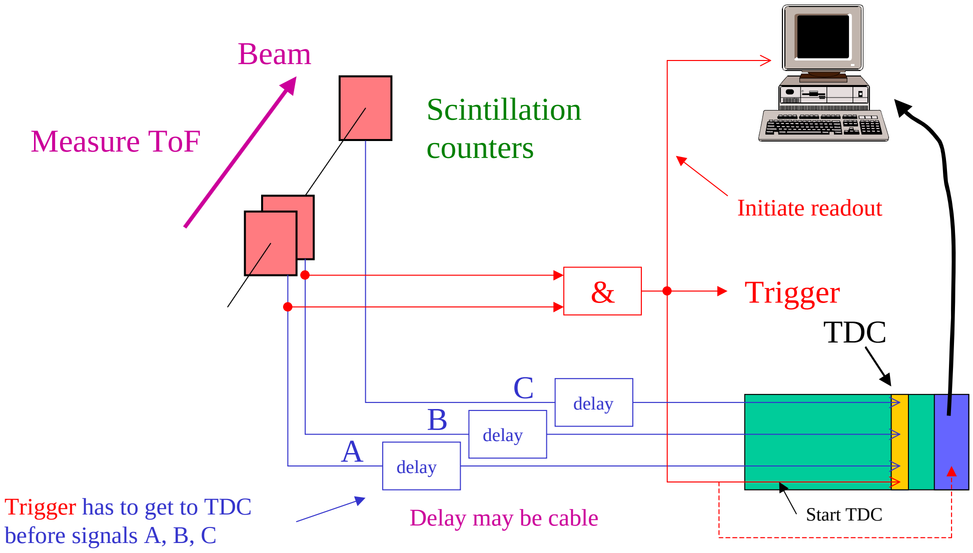


Fig. 1: Example of a simple experiment with its T/DAQ system

The 'readout model' of this traditional approach to T/DAQ is illustrated in Fig. 2, which shows the sequence of actions — arrival of the trigger, arrival of the detector signals (followed by digitization and storage in a data register in the TDC), and readout into the DAQ computer. Since no new trigger can be accepted until the readout is complete, the readout dead-time is given by the product of the trigger rate and the readout time.

[2] NIM and CAMAC modules are electronic modules that conform to agreed standards — modules for many functions needed in a T/DAQ system are available commercially.

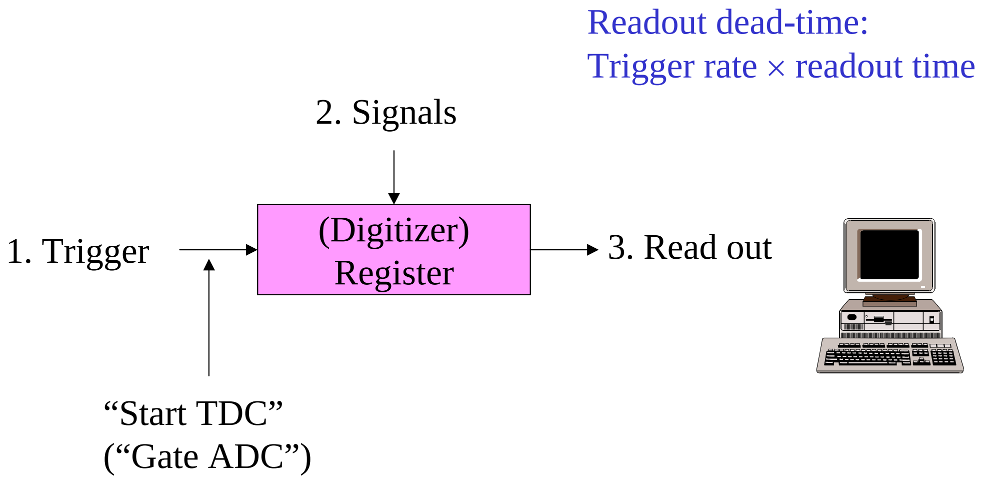


Fig. 2: Readout model in the 'traditional' approach

2.2 Local buffer

The traditional approach described above can be improved by adding a local 'buffer' memory into which the data are moved rapidly following a trigger, as illustrated in Fig. 3. This fast readout reduces the dead-time, which is now given by the product of the trigger rate and the local readout time. This approach is particularly useful in large systems, where the transfer of data can proceed in parallel with many local buffers (e.g., one local buffer for each crate of electronics) — local readout can remain fast even in a large system. Also, the data may be moved more quickly into the local buffer within the crate than into the DAQ computer. Note that the dashed line in the bottom, right-hand part of Fig. 1 indicates this extension to the traditional approach — the trigger signal is used to initiate the local readout within the crate.

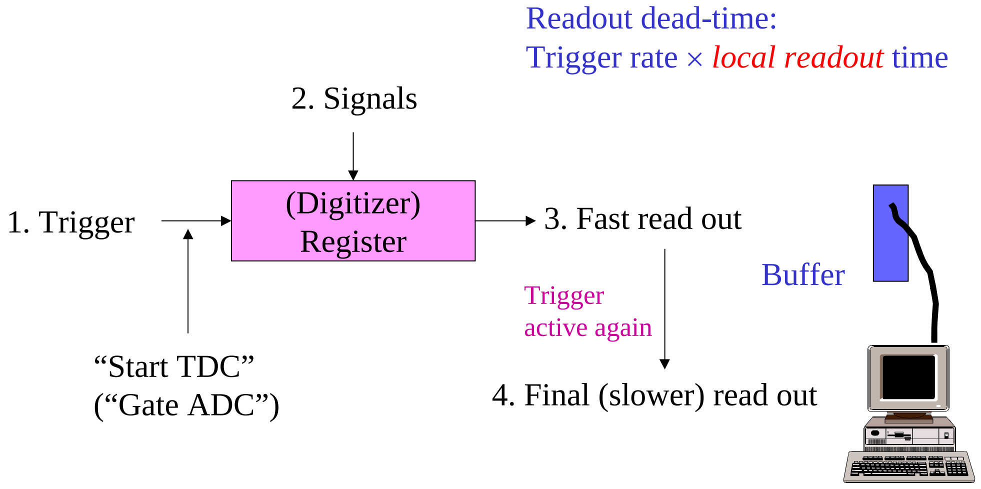


Fig. 3: Readout system with local buffer memory

The addition of a local buffer reduces the effective readout time, but the requirement of a fast trigger still remains. Signals have to be delayed until the trigger decision is available at the digitizers. This is not easy to achieve, even with very simple trigger logic — typically one relies on using fast (air-core) cables for trigger signals with the shortest possible routing so that the trigger signals arrive before the rest of the signals (which follow a longer routing on slower cables). It is not possible to apply complex selection criteria on this time-scale.

2.3 Multi-level triggers

It is not always possible to simultaneously meet the physics requirements (high efficiency, high background rejection) and achieve an extremely short trigger 'latency' (time to form the trigger decision and distribute it to the digitizers). A solution is to introduce the concept of multi-level triggers, where the first level has a short latency and maintains high efficiency, but only has a modest rejection power. Further background rejection comes from the higher trigger levels that can be slower. Sometimes the very fast first stage of the trigger is called the 'pre-trigger' — it may be sufficient to signal the presence of minimal activity in the detectors at this stage.

The use of a pre-trigger is illustrated in Fig. 4. Here the pre-trigger is used to provide the start signal to the TDCs (and to gate ADCs, etc.), while the main trigger (which can come later and can therefore be based on more complex calculations) is used to initiate the readout. In cases where the pre-trigger is not confirmed by the main trigger, a 'fast clear' is used to re-activate the digitizers (TDCs, ADCs, etc.).

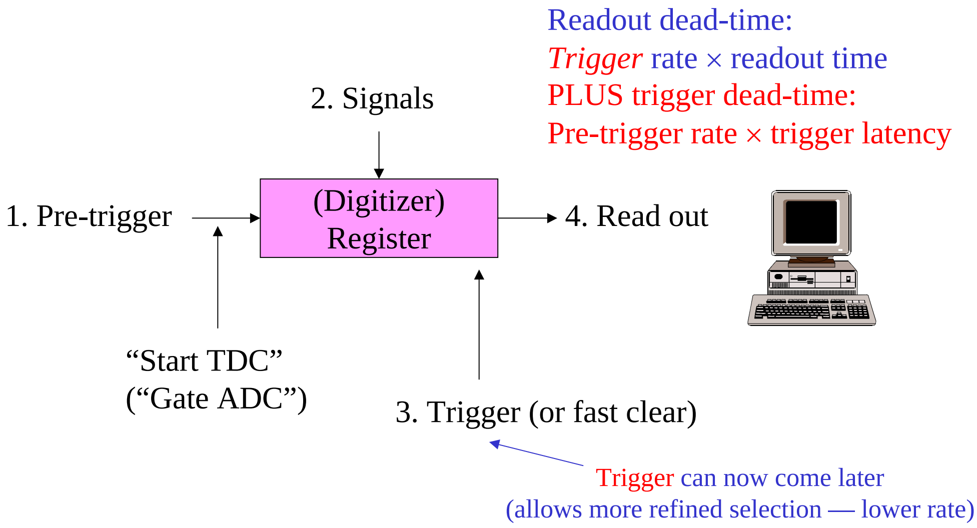

Readout dead-time:
Trigger rate × readout time
PLUS trigger dead-time:
Pre-trigger rate × trigger latency
2. Signals
1. Pre-trigger
(Digitizer)
Register
4. Read out
"Start TDC"
("Gate ADC")
3. Trigger (or fast clear)
Trigger can now come later
(allows more refined selection — lower rate)

Fig. 4: Readout system with pre-trigger and fast clear

Using a pre-trigger (but without using a local buffer for now), the dead-time has two components. Following each pre-trigger there is a dead period until the trigger or fast clear is issued (defined here as the trigger latency). For the subset of pre-triggers that give rise to a trigger, there is an additional dead period given by the readout time. Hence, the total dead-time is given by the product of the pre-trigger rate and the trigger latency, added to the product of the trigger rate and the readout time.

The two ingredients — use of a local buffer and use of a pre-trigger with fast clear — can be combined as shown in Fig. 5, further reducing the dead-time. Here the total dead-time is given by the product of the pre-trigger rate and the trigger latency, added to the product of the trigger rate and the local readout time.

2.4 Further improvements

The idea of multi-level triggers can be extended beyond having two levels (pre-trigger and main trigger). One can have a series of trigger levels that progressively reduce the rate. The efficiency for the desired physics must be kept high at all levels since rejected events are lost forever. The initial levels can have modest rejection power, but they must be fast since they see a high input rate. The final levels must have strong rejection power, but they can be slower because they see a much lower rate (thanks to the rejection from the earlier levels).

In a multi-level trigger system, the total dead-time can be written as the sum of two parts: the trigger dead-time summed over trigger levels, and the readout dead-time. For a system with N levels, this can be written

$$(\sum_{i=2}^{N} R_{i-1} \times L_i) + R_N \times T_{\mathrm{LRO}}$$

where R_i is the rate after the i^{th} trigger level, L_i is the latency of the i^{th} trigger level, and T_{LRO} is the local readout time. Note that R_1 corresponds to the pre-trigger rate.

In the above, two implicit assumptions have been made: (1) that all trigger levels are completed before the readout starts, and (2) that the pre-trigger (i.e., the lowest-level trigger) is available by the time the first signals from the detector arrive at the digitizers. The first assumption results in a long dead period for some events — those that survive the first (fast) levels of selection. The dead-time can be reduced by moving the data into intermediate storage after the initial stages of trigger selection, after which further low-level triggers can be accepted (in parallel with the execution of the later stages of trigger selection on the first event). The second assumption can also be avoided, e.g., in collider experiments with bunched beams as discussed below.

In the next section, aspects of particle colliders that affect the design of T/DAQ systems are introduced. Afterwards, the discussion returns to readout models and dead-time, considering the example of LEP experiments.

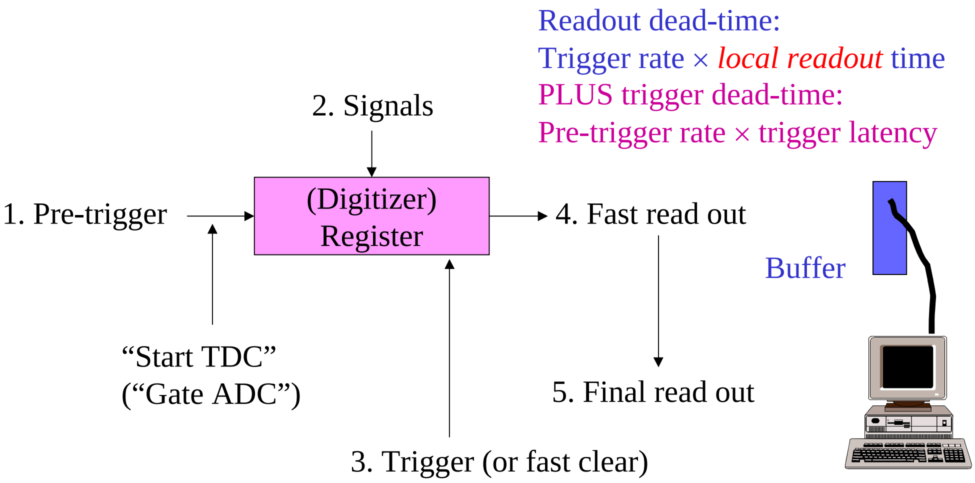


Fig. 5: Readout system using both pre-trigger and local buffer

3 Collider experiments

In high-energy particle colliders (HERA, LEP, LHC, Tevatron), the particles in the counter-rotating beams are bunched. Bunches of particles cross at regular intervals and interactions only occur during the bunch crossings. Here the trigger has the job of selecting the *bunch crossings* of interest for physics analysis, i.e., those containing interactions of interest.

In the following notes, the term 'event' is used to refer to the record of all the products from a given bunch crossing (plus any activity from other bunch crossings that gets recorded along with this). Be aware (and beware!) — the term 'event' is not uniquely defined! Some people use the term 'event' for the products of a single interaction between incident particles. Note that many people use 'event' interchangeably to mean different things.

In $e^+ e^-$ colliders, the interaction rate is very small compared to the bunch-crossing rate (because of the low $e^+ e^-$ cross-section). Generally, selected events contain just one interaction — i.e., the event

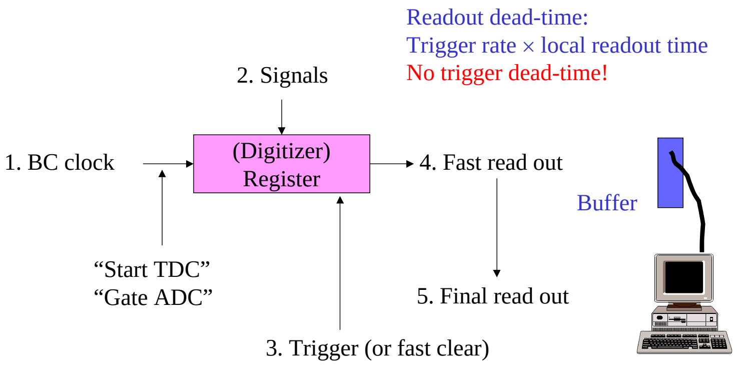


Fig. 6: Readout system using bunch-crossing (BC) clock and fast clear

is generally a single interaction. This was the case at LEP and is also true at the e–p collider HERA. In contrast, at LHC with the design luminosity $\mathscr{L}$ of $10^{24}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, each bunch crossing will contain on average about 25 interactions as discussed below. This means that an interaction of interest, e.g., one that produced $\mathrm{H} \rightarrow \mathrm{ZZ} \rightarrow \mathrm{e}^+\mathrm{e}^-\mathrm{e}^+\mathrm{e}^-$, will be recorded together with 25 other proton–proton interactions that occurred in the same bunch crossing. The interactions that make up the 'underlying event' are often called 'minimum-bias' interactions because they are the ones that would be selected by a trigger that selects interactions in an unbiased way. The presence of additional interactions that are recorded together with the one of interest is sometimes referred to as 'pile-up'.

A further complication is that particle detectors do not have an infinitely fast response time — this is analogous to the exposure time of a camera. If the 'exposure time' is shorter than the bunch-crossing period, the event will contain only information from the selected bunch crossing. Otherwise, the event will contain, in addition,any activity from neighbouring bunches. In $\mathrm{e}^+\,\mathrm{e}^-$ colliders (e.g., LEP) it is very unlikely for there to be any activity in nearby bunch crossings, which allows the use of slow detectors such as the time projection chamber (TPC). This is also true at HERA and in the ALICE experiment [1] at LHC that will study heavy-ion collisions at much lower luminosities than in the proton–proton case.

The bunch-crossing period for proton–proton collisions at LHC will be only 25 ns (a 40 MHz rate). At the design luminosity the interaction rate will be $\mathcal{O}(10^9)$ Hz and, even with the short bunch-crossing period, there will be an average of about 25 interactions per bunch crossing. Some detectors, for example the ATLAS silicon tracker, achieve an exposure time of less than 25 ns, but many do not. For example, pulses from the ATLAS liquid-argon calorimeter extend over many bunch crossings.

4 Design of a trigger and data-acquisition system for LEP

Let us now return to the discussion of designing a T/DAQ system, considering the case of experiments at LEP (ALEPH [2], DELPHI [3], L3 [4], and OPAL [5]), and building on the model developed in Section 2.

4.1 Using the bunch-crossing signal as a 'pre-trigger'

If the time between bunch crossings (BCs) is reasonably long, one can use the clock that signals when bunches of particles cross as the pre-trigger. The first-level trigger can then use the time between bunch crossings to make a decision, as shown in Fig. 6. For most crossings the trigger will reject the event by issuing a fast clear — in such cases no dead-time is introduced. Following an 'accept' signal, dead-time will be introduced until the data have been read out (or until the event has been rejected by a higher-level trigger). This is the basis of the model that was used at LEP, where the bunch-crossing interval of 22 µs (11 µs in eight-bunch mode) allowed comparatively complicated trigger processing (latency of a few microseconds). Note that there is no first-level trigger dead-time because the decision is made during the

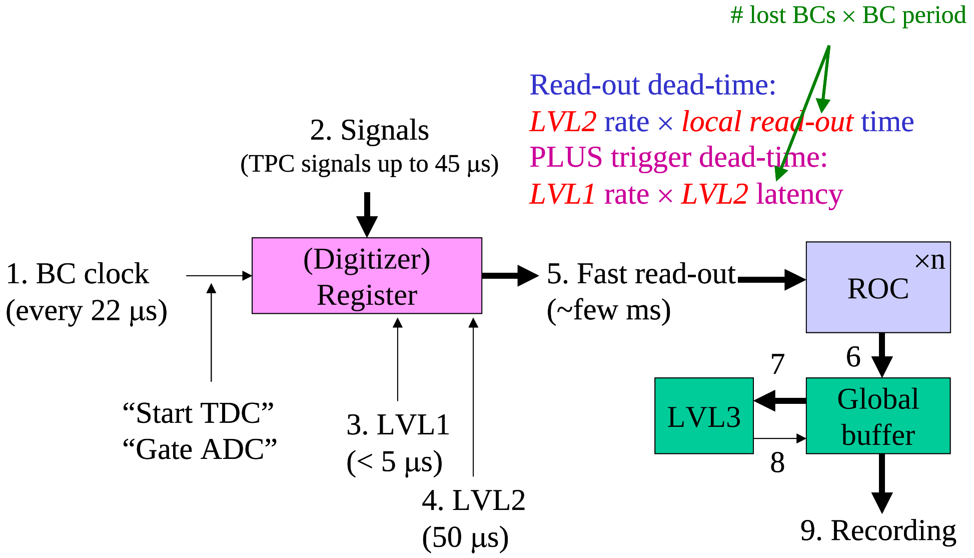


Fig. 7: LEP readout model (ALEPH)

interval between bunch crossings where no interactions occur. As discussed below, the trigger rates were reasonably low (very much less than the BC rate), giving acceptable dead-time due to the second-level trigger latency and the readout.

In the following, the readout model used at LEP is illustrated by concentrating on the example of ALEPH [2][3]. Figure 7 shows the readout model, using the same kind of block diagram as presented in Section 2. The BC clock is used to start the TDCs and generate the gate for the ADCs, and a first-level (LVL1) trigger decision arrives in less than 5 μs so that the fast clear can be completed prior to the next bunch crossing. For events retained by LVL1, a more sophisticated second-level (LVL2) trigger decision is made after a total of about 50 μs. Events retained by LVL2 are read out to local buffer memory (within the readout controllers or 'ROCs'), and then passed to a global buffer. There is a final level of selection (LVL3) before recording the data on permanent storage for offline analysis.

For readout systems of the type shown in Fig. 7, the total dead-time is given by the sum of two components — the trigger dead-time and the readout dead-time:

The trigger dead-time is evaluated by counting the number of BCs that are lost following each LVL1 trigger, then calculating the product of the LVL1 trigger rate, the number of lost BCs and the BC period. Note that the effective LVL2 latency, given by the number of lost BCs and the BC period, is less than (or equal to) the true LVL2 latency.

The readout dead-time is given by the product of the LVL2 trigger rate and the time taken to perform local readout into the ROCs. Strictly speaking, one should also express this dead-time in terms of the number of BCs lost after the LVL2 trigger, but since the readout time is much longer than the BC period the difference is unimportant. Note that, as long as the buffers in the ROCs and the global buffers do not fill up, no additional dead-time is introduced by the final readout and the LVL3 trigger.

Let us now look quantitatively at the example of the DELPHI experiment for which the T/DAQ system was similar to that described above for ALEPH. Typical numbers for LEP-II are shown in Table 1 [3].

[3] The author was not involved in any of the LEP experiments. In these lectures the example of ALEPH is used to illustrate how triggers and data acquisition were implemented at LEP; some numbers from DELPHI are also presented. The T/DAQ systems in all of the LEP experiments were conceptually similar.

Table 1: Typical T/DAQ parameters for the DELPHI experiment at LEP-II

Quantity	**Value**
LVL1 rate	~ 500–1000 Hz (instrumental background)
LVL2 rate	6–8 Hz
LVL3 rate	4–6 Hz
LVL2 latency	38 μs (1 lost BC ⇒ 22 μs effective)
Local readout time	~ 2.5 ms
Readout dead-time	~ 7 Hz × 2.5 · 10^{-3} s = 1.8%
Trigger dead-time	~ 750 Hz × 22 · 10^{-6} s = 1.7%
Total dead-time	~ 3–4%

4.2 Data acquisition at LEP

Let us now continue our examination of the example of the ALEPH T/DAQ system. Following a LVL2 trigger, events were read out locally and in parallel within the many readout crates — once the data had been transferred within each crate to the ROC, further LVL1 and LVL2 triggers could be accepted. Subsequently, the data from the readout crates were collected by the main readout computer, 'building' a complete event. As shown in Fig. 8, event building was performed in two stages: an event contained a number of sub-events, each of which was composed of several ROC data blocks. Once a complete event was in the main readout computer, the LVL3 trigger made a final selection before the data were recorded.

The DAQ system used a hierarchy of computers — the local ROCs in each crate; event builders (EBs) for sub-events; the main EB; the main readout computer. The ROCs performed some data processing (e.g., applying calibration algorithms to convert ADC values to energies) in addition to reading out the data from ADCs, TDCs, etc. (Zero suppression was already performed at the level of the digitizers where appropriate.) The first layer of EBs combined data read out from the ROCs of individual sub-detectors into sub-events; then the main EB combined the sub-events for the different sub-detectors. Finally, the main readout computer received full events from the main EB, performed the LVL3 trigger selection, and recorded selected events for subsequent analysis.

As indicated in Fig. 9, event building was bus based — each ROC collected data over a bus from the digitizing electronics; each sub-detector EB collected data from several ROCs over a bus; the main EB collected data from the sub-detector EBs over a bus. As a consequence, the main EB and the main readout computer saw the full data rate prior to the final LVL3 selection. At LEP this was fine — with an event rate after LVL2 of a few hertz and an event size of 100 kbytes, the data rate was a few hundred kilobytes per second, much less than the available bandwidth (e.g., 40 Mbytes/s maximum on VME bus).

4.3 Triggers at LEP

The triggers at LEP aimed to select any $e^+ e^-$ annihilation event with a visible final state, including events with little visible energy, plus some fraction of two-photon events, plus Bhabha scattering events. Furthermore, they aimed to select most events by multiple, independent signatures so as to maximize the trigger efficiency and to allow the measurement of the efficiency from the data. The probability for an event to pass trigger A or trigger B is ~ $1 - \delta_A \delta_B$, where δ_A and δ_B are the individual trigger inefficiencies, which is very close to unity for small δ. Starting from a sample of events selected with trigger A, the efficiency of trigger B can be estimated as the fraction of events passing trigger B in addition. Note that in the actual calculations small corrections were applied for correlations between the trigger efficiencies.

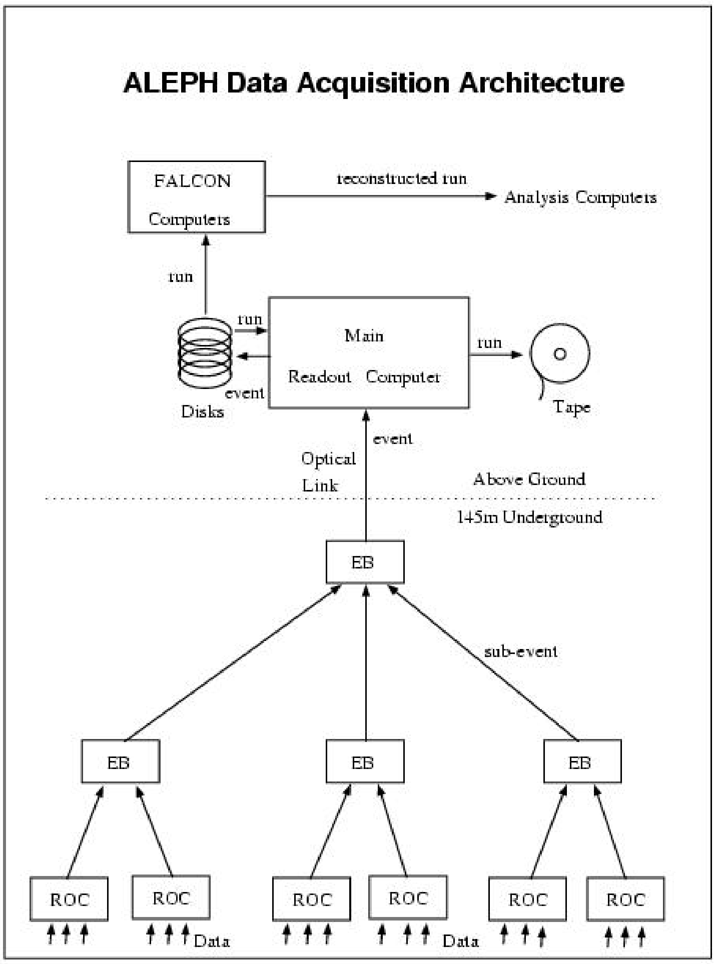


Fig. 8: ALEPH data-acquisition architecture

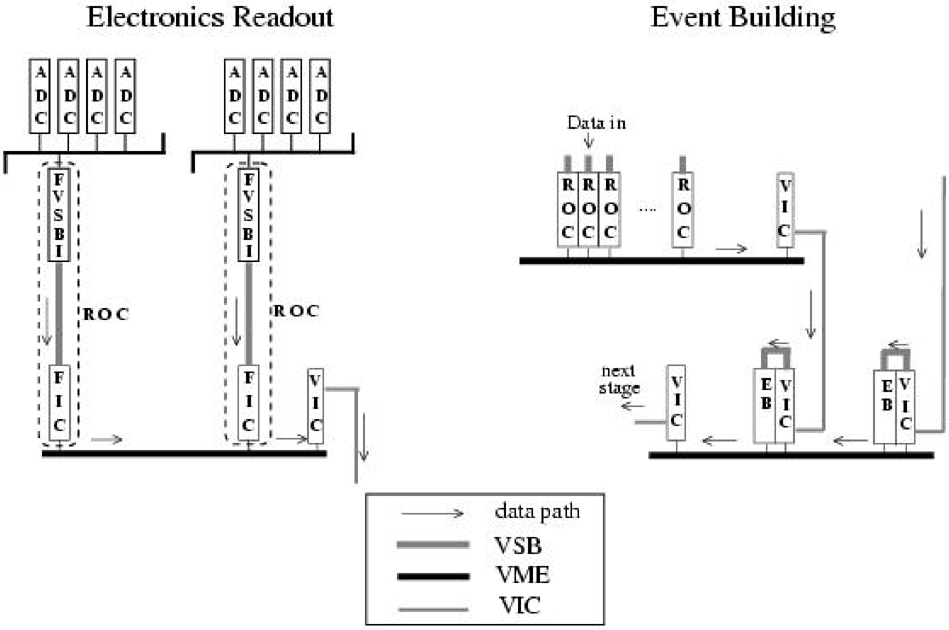


Fig. 9: Event building in ALEPH

5 Physics requirements — two examples

In the following, the physics requirements on the T/DAQ systems at LEP and at LHC are examined. These are complementary cases — at LEP precision physics was the main emphasis, at LHC discovery physics will be the main issue. Precision physics at LEP needed accurate determination of the absolute cross-section (e.g., in the determination of the number of light-neutrino species). Discovery physics at LHC will require sensitivity to a huge range of predicted processes with diverse signatures (with very low signal rates expected in some cases), aiming to be as sensitive as possible to new physics that has not been predicted (by using inclusive signatures). This has to be achieved in the presence of an enormous rate of Standard Model physics backgrounds (the rate of proton–proton collisions at LHC will be $\mathcal{O}(10^9)$ Hz — $\sigma \sim 100$ mb, $\mathscr{L} \sim 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$).

5.1 Physics requirements at LEP

Triggers at LEP aimed to identify all events coming from $\mathrm{e}^+\ \mathrm{e}^-$ annihilations with visible final states. At LEP-I, operating with $\sqrt{s} \sim m_{\mathrm{Z}}$, this included $\mathrm{Z} \rightarrow$ hadrons, $\mathrm{Z} \rightarrow \mathrm{e}^+\mathrm{e}^-$, $\mathrm{Z} \rightarrow \mu^+\mu^-$, and $\mathrm{Z} \rightarrow \tau^+\tau^-$; at LEP-II, operating above the W W threshold, this included W W, Z Z and single-boson events. Sensitivity was required even in cases where there was little visible energy, e.g., in the Standard Model for $\mathrm{e}^+\mathrm{e}^- \rightarrow \mathrm{Z}\gamma$, with $\mathrm{Z} \rightarrow \nu\nu$, and in new-particle searches such as $\mathrm{e}^+\mathrm{e}^- \rightarrow \tilde{\chi}^+\tilde{\chi}^-$ for the case of small $\tilde{\chi}^\pm\tilde{\chi}^0$ mass difference that gives only low-energy visible particles ($\tilde{\chi}^0$ is the lightest supersymmetric particle). In addition, the triggers had to retain some fraction of two-photon collision events (used for QCD studies), and identify Bhabha scatters (needed for precise luminosity determination).

The triggers could retain events with any significant activity in the detector. Even when running at the Z peak, the rate of Z decays was only $\mathcal{O}(1)$ Hz — physics rate was not an issue. The challenge was in maximizing the efficiency (and acceptance) of the trigger, and making sure that the small inefficiencies were very well understood. The determination of absolute cross-section depends on knowing the integrated luminosity and the experimental efficiency to select the process in question (i.e., the efficiency to trigger on the specific physics process). Precise determination of the integrated luminosity required excellent understanding of the trigger efficiency for Bhabha-scattering events (luminosity determined from the rate of Bhabha scatters within a given angular range). A major achievement at LEP was to reach 'per mil' precision.

The trigger rates (events per second) and the DAQ rates (bytes per second) at LEP were modest as, discussed in Section 4.

5.2 Physics requirements at the LHC

Triggers in the general-purpose proton–proton experiments at the LHC (ATLAS [6] and CMS [7]) will have to retain as high as possible a fraction of the events of interest for the diverse physics programmes of these experiments. Higgs searches in and beyond the Standard Model will include looking for $\mathrm{H} \rightarrow \mathrm{ZZ} \rightarrow$ leptons and also $\mathrm{H} \rightarrow \mathrm{b}\overline{\mathrm{b}}$. Supersymmetry (SUSY) searches will be performed with and without the assumption of R-parity conservation. One will search for other new physics using inclusive triggers that one hopes will be sensitive to unpredicted processes. In parallel with the searches for new physics, the LHC experiments aim to do precision physics, such as measuring the W mass and some B-physics studies, especially in the early phases of LHC running when the luminosity is expected to be comparatively low.

In contrast to the experiments at LEP, the LHC trigger systems have a hard job to reduce the physics event rate to a manageable level for data recording and offline analysis. As discussed above, the design luminosity $\mathscr{L} \sim 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, together with $\sigma \sim 100$ mb, implies an $\mathcal{O}(10^9)$ Hz interaction rate. Even the rate of events containing leptonic decays of W and Z bosons is $\mathcal{O}(100)$ Hz. Furthermore, the size of the events is very large, $\mathcal{O}(1)$ Mbyte, reflecting the huge number of detector channels and the high particle multiplicity in each event. Recording and subsequently processing offline $\mathcal{O}(100)$ Hz event rate per experiment with an $\mathcal{O}(1)$ Mbyte event size is considered feasible, but it implies major computing resources [8]. Hence, only a tiny fraction of proton–proton collisions can be selected — taking the order-of-magnitude numbers given above, the maximum fraction of interactions that can be selected is $\mathcal{O}(10^{-7})$. Note that the general-purpose LHC experiments have to balance the needs of maximizing physics coverage and reaching acceptable (i.e., affordable) recording rates.

The LHCb experiment [9], which is dedicated to studying B-physics, faces similar challenges to ATLAS and CMS. It will operate at a comparatively low luminosity ($\mathscr{L} \sim 10^{32}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$), giving an overall proton–proton interaction rate of ~ 20 MHz — chosen to maximize the rate of single-interaction bunch crossings. The event size will be comparatively small (~ 100 kbytes) as a result of having fewer detector channels and of the lower occupancy of the detector (due to the lower luminosity with less pile-up). However, there will be a very high rate of beauty production in LHCb — taking $\sigma \sim$ μb, the production rate will be ~ 100 kHz — and the trigger must search for specific B-decay modes that are of interest for physics analysis, with the aim of recording an event rate of only ~ 200 Hz.

The heavy-ion experiment ALICE [1] is also very demanding, particularly from the DAQ point of view. The total interaction rate will be much smaller than in the proton–proton experiments — $\mathscr{L} \sim 10^{27}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1} \Rightarrow$ rate ~ 8000 Hz for Pb–Pb collisions. However, the event size will be huge due to the high final-state multiplicity in Pb–Pb interactions at LHC energy. Up to $\mathcal{O}(10^4)$ charged particles will be produced in the central region, giving an event size of up to ~ 40 Mbytes when the full detector is read out. The ALICE trigger will select 'minimum-bias' and 'central' events (rates scaled down to a total of about 40 Hz), and events with dileptons (~ 1 kHz with only part of the detector read out). Even compared to the other LHC experiments, the volume of data to be stored and subsequently processed offline will be massive, with a data rate to storage of ~ 1 Gbytes/s (considered to be about the maximum affordable rate).

6 Signatures of different types of particle

The generic signatures for different types of particle are illustrated in Fig. 10; much more detail can be found in Ref. [10]. Moving away from the interaction point (shown as a star on the left-hand side of Fig. 10), one finds the inner tracking detector (IDET), the electromagnetic calorimeter (ECAL), the hadronic calorimeter (HCAL) and the muon detectors (MuDET). Charged particles (electrons, muons and charged hadrons) leave tracks in the IDET. Electrons and photons shower in the ECAL, giving localized clusters of energy without activity in the HCAL. Hadrons produce larger showers that may start in the ECAL but extend into the HCAL. Muons traverse the calorimeters with minimal energy loss and are detected in the MuDET.

The momenta of charged particles are measured from the radii of curvature of their tracks in the IDET which is embedded in a magnetic field. A further measurement of the momenta of muons may be made in the MuDET using a second magnet system. The energies of electrons, photons and hadrons are measured in the calorimeters. Although neutrinos leave the detector system without interaction, one can infer their presence from the momentum imbalance in the event (sometimes referred to as 'missing energy'). Hadronic jets contain a mixture of particles, including neutral pions that decay almost immediately into photon pairs that are then detected in the ECAL. The jets appear as broad clusters of energy in the calorimeters where the individual particles will sometimes not be resolved.

7 Selection criteria and trigger implementations at LEP

The details of the selection criteria and trigger implementations at LEP varied from experiment to experiment [2–5]. Discussion of the example of ALEPH is continued with the aim of giving a reasonably in-depth view of one system. For triggering purposes, the detector was divided into segments with a total of 60 regions in θ, ϕ (θ is polar angle and ϕ is azimuth with respect to the beam axis). Within these segments, the following trigger objects were identified:

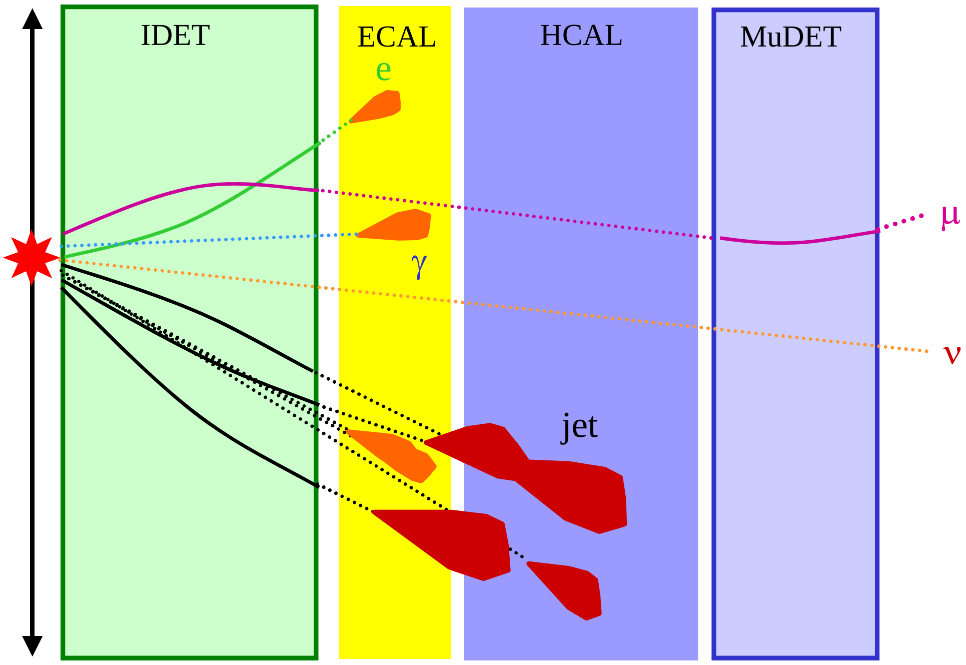


Fig. 10: Signatures of different types of particle in a generic detector

1. muon — requiring a track penetrating the hadron calorimeter and seen in the inner tracker;
2. charged electromagnetic (EM) energy — requiring an EM calorimeter cluster and a track in the inner tracker;
3. neutral EM energy — requiring an EM calorimeter cluster (with higher thresholds than in 2. to limit the rate to acceptable levels).

In addition to the above local triggers, there were total-energy triggers (applying thresholds on energies summed over large regions — the barrel or a full endcap), a back-to-back tracks trigger, and triggers for Bhabha scattering (luminosity monitor).

The LVL1 triggers were implemented using a combination of analog and digital electronics. The calorimeter triggers were implemented using analogue electronics to sum signals before applying thresholds on the sums. The LVL1 tracking trigger looked for patterns of hits in the inner-tracking chamber (ITC) consistent with a track with $p_{\mathrm{T}} > 1$ GeV [4] — at LVL2 the TPC was used instead. The final decision was made by combining digital information from calorimeter and tracking triggers, making local combinations within segments of the detector, and then making a global combination (logical OR of conditions).

8 Towards the LHC

In some experiments it is not practical to make a trigger in the time between bunch crossings because of the short BC period — the BC interval is 132 ns at Tevatron-II, 96 ns at HERA and 25 ns at LHC. In such cases the concept of ‘pipelined’ readout has to be introduced (also pipelined LVL1 trigger processing). Furthermore, in experiments at high-luminosity hadron colliders the data rates after the LVL1 trigger selection are very high, and new ideas have to be introduced for the high-level triggers (HLTs) and DAQ — in particular, event building has to be based on data networks and switches rather than data buses.

[4] Here, p_{T} is transverse momentum (measured with respect to the beam axis); similarly, E_{T} is transverse energy.

8.1 Pipelined readout

In pipelined readout systems (see Fig. 11), the information from each BC, for each detector element, is retained during the latency of the LVL1 trigger (several µs). The information may be retained in several forms — analog levels (held on capacitors); digital values (e.g., ADC results); binary values (i.e., hit or no hit). This is done using a logical 'pipeline', which may be implemented using a first-in, first-out (FIFO) memory circuit. Data reaching the end of the pipeline are either discarded or, in the case of a trigger accept decision, moved to a secondary buffer memory (small fraction of BCs).

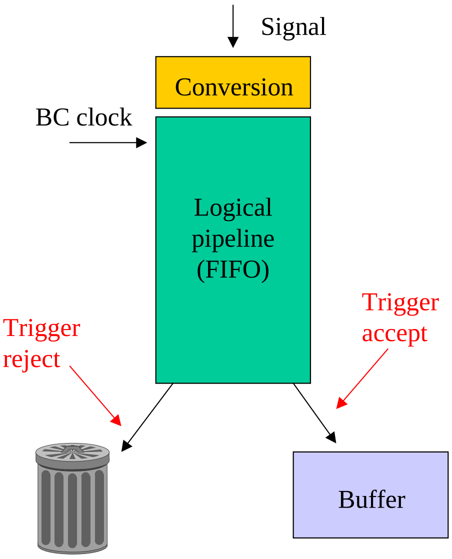


Fig. 11: Example of pipelined readout

Pipelined readout systems will be used in the LHC experiments (they are already being used in existing experiments at HERA [11, 12] and the Tevatron [13, 14], but the demands at LHC are even greater because of the short BC period). A typical LHC pipelined readout system is illustrated in Fig. 12, where the digitizer and pipeline are driven by the 40 MHz BC clock. A LVL1 trigger decision is made for each bunch crossing (i.e., every 25 ns), although the LVL1 latency is several microseconds — the LVL1 trigger must concurrently process many events (this is achieved by using pipelined trigger processing as discussed below).

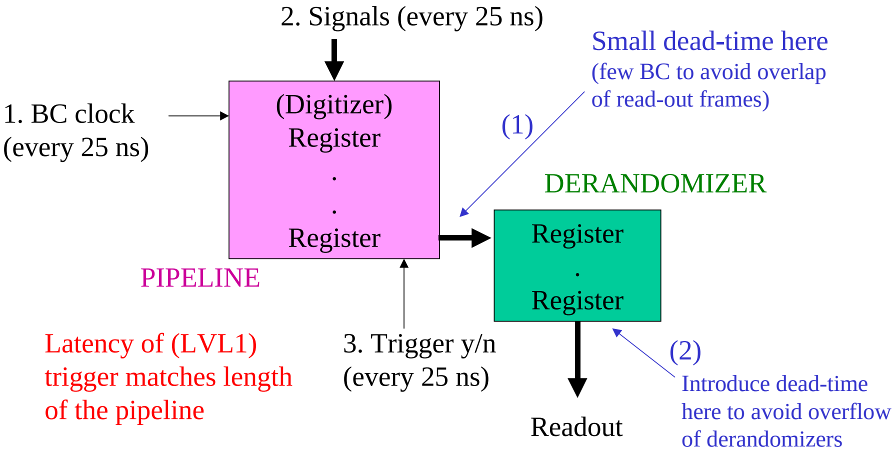

2. Signals (every 25 ns)
Small dead-time here
(few BC to avoid overlap of read-out frames)
1. BC clock (every 25 ns)
(Digitizer) Register
.
.
Register
(1)
DERANDOMIZER
Register
.
Register
PIPELINE
Latency of (LVL1) trigger matches length of the pipeline
3. Trigger y/n (every 25 ns)
(2)
Introduce dead-time here to avoid overflow of derandomizers
Readout

Fig. 12: Pipelined readout with derandomizer at LHC

The data for events that are selected by the LVL1 trigger are transferred into a 'derandomizer' — a memory that can accept the high instantaneous input rate (i.e., one word per 25 ns) while being read out at the much lower average data rate (determined by the LVL1 trigger rate rather than the BC rate). In principle no dead-time needs to be introduced in such a system. However, in practice, data are retained for a few BCs around the one that gave rise to the trigger, and a dead period of a few BCs is introduced to ensure that the same data do not have to be accessed for more than one trigger. Dead-time must also be introduced to prevent the derandomizers from overflowing, e.g., where, due to a statistical fluctuation, many LVL1 triggers arrive in quick succession. The dead-time from the first of these sources can be estimated as follows (numbers from ATLAS): taking a LVL1 trigger rate of 75 kHz and 4 dead BCs following each LVL1 trigger gives 75 kHz $\times$ 4 $\times$ 25 ns $=$ 0.75%. The dead-time from the second source depends on the size of the derandomizer and the speed with which it can be emptied — in ATLAS the requirements are $< 1\%$ dead-time for a LVL1 rate of 75 kHz ($< 6\%$ for 100 kHz).

Some of the elements of the readout chain in the LHC experiments have to be mounted on the detectors (and hence are totally inaccessible during running of the machine and are in an environment with high radiation levels). This is shown for the case of CMS in Fig. 13.

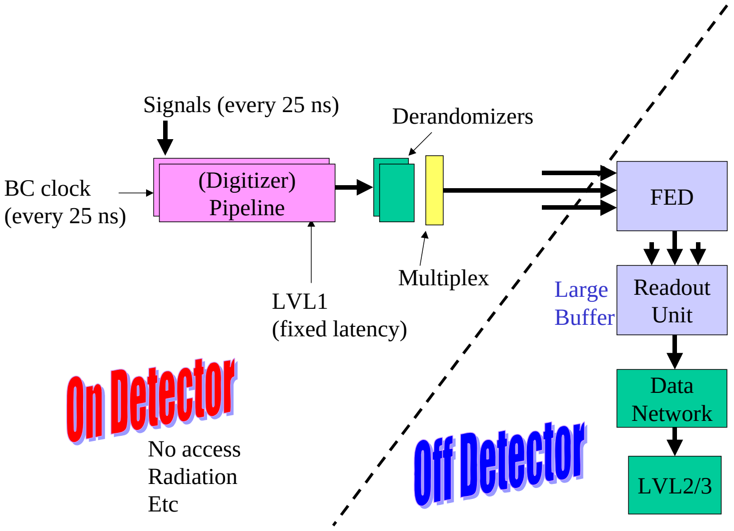


Fig. 13: Location of readout components in CMS

There are a variety of options for the placement of digitization in the readout chain, and the optimum choice depends on the characteristics of the detector in question. Digitization may be performed on the detector at 40 MHz rate, prior to a digital pipeline (e.g., CMS calorimeter). Alternatively, it may be done on the detector after multiplexing signals from several analog pipelines (e.g., ATLAS EM calorimeter) — here the digitization rate can be lower, given by the LVL1 trigger rate multiplied by the number of signals to be digitized per trigger. Another alternative (e.g., CMS tracker) is to multiplex analog signals from the pipelines over analog links, and then to perform the digitization off-detector.

8.2 Pipelined LVL1 trigger

As discussed above, the LVL1 trigger has to deliver a new decision every BC, although the trigger latency is much longer than the BC period; the LVL1 trigger must concurrently process many events. This can be achieved by 'pipelining' the processing in custom trigger processors built using modern digital electronics. The key ingredients in this approach are to break the processing down into a series of steps, each of which can be performed within a single BC period, and to perform many operations in parallel by having separate processing logic for each calculation. Note that in such a system the latency of the LVL1 trigger is fixed — it is determined by the number of steps in the calculation, plus the time taken to move signals and data to, from and between the components of the trigger system (e.g., propagation delays on cables).

Pipelined trigger processing is illustrated in Fig. 14 — as will be seen later, this example corresponds to a (very small) part of the ATLAS LVL1 calorimeter trigger processor. The drawing on the left of Fig. 14 depicts the EM calorimeter as a grid of 'towers' in η–ϕ space (η is pseudorapidity, ϕ is azimuth angle). The logic shown on the right determines if the energy deposited in a horizontal or vertical pair of towers in the region [A, B, C] exceeds a threshold. In each 25 ns period, data from one layer of 'latches' (memory registers) are processed through the next step in the processing 'pipe', and the results are captured in the next layer of latches. Note that, in the real system, such logic has to be performed in parallel for ~ 3500 positions of the reference tower; the tower 'A' could be at any position in the calorimeter. In practice, modern electronics is capable of doing more than a simple add or compare operation in 25 ns, so there is more logic between the latches than in this illustration.

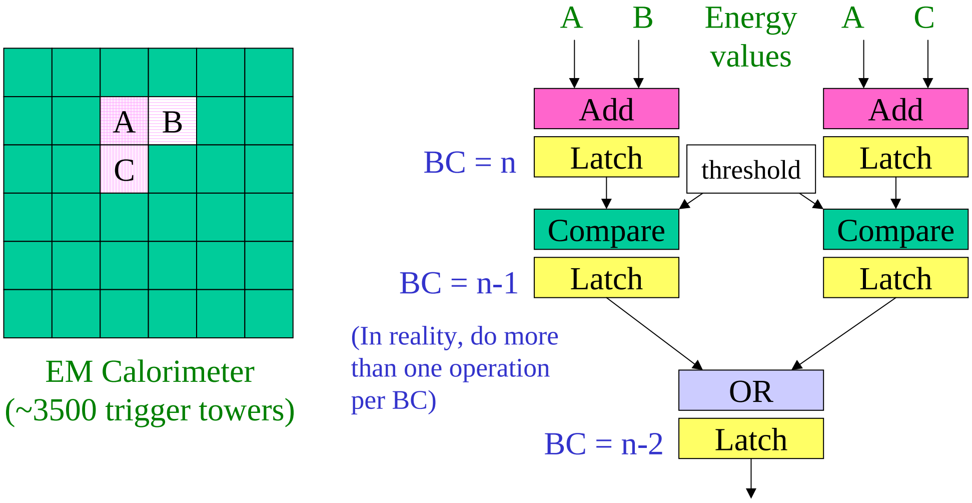


Fig. 14: Illustration of pipelined processing

The amount of data to be handled varies with depth in the processing pipeline, as indicated in Fig. 15. Initially the amount of data expands compared to the raw digitization level since each datum typically participates in several operations — the input data need to be 'fanned out' to several processing elements. Subsequently the amount of data decreases as one moves further down the processing tree. The final trigger decision can be represented by a single bit of information for each BC — yes or no (binary 1 or 0). Note that, in addition to the trigger decision, the LVL1 processors produce a lot of data for use in monitoring the system and to guide the higher levels of selection.

Although they have not been discussed in these lectures because of time limitations, some fixed-target experiments have very challenging T/DAQ requirements. Some examples can be found in Refs. [15, 16].

9 Selection criteria at LHC

Features that distinguish new physics from the bulk of the cross-section for Standard Model processes at hadron colliders are generally the presence of high-p_T particles (or jets). For example, these may be the products of the decays of new heavy particles. In contrast, most of the particles produced in minimum-bias interactions are soft ($p_T \sim 1$ GeV or less). More specific signatures are the presence of high-p_T leptons (e, μ, τ), photons and/or neutrinos. For example, these may be the products (directly or indirectly) of new heavy particles. Charged leptons, photons and neutrinos give a particularly clean signature (c.f. low-p_T hadrons in minimum-bias events), especially if they are 'isolated' (i.e., not inside jets). The presence of heavy particles such as W and Z bosons can be another signature for new

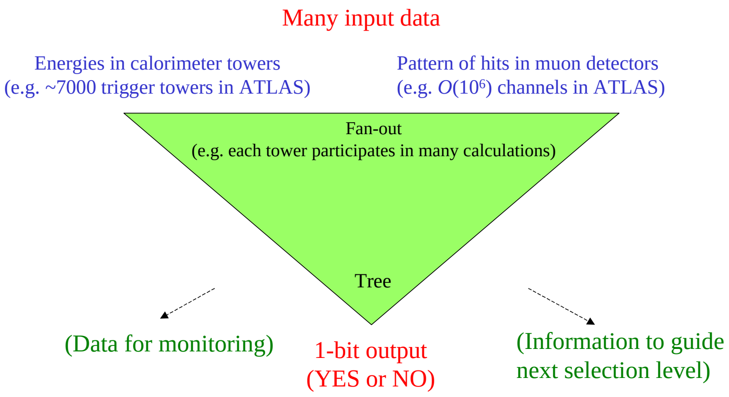

Many input data
Energies in calorimeter towers
(e.g. ~7000 trigger towers in ATLAS)
Pattern of hits in muon detectors
(e.g. $O(10^6)$ channels in ATLAS)
Fan-out
(e.g. each tower participates in many calculations)
Tree
(Data for monitoring)
1-bit output
(YES or NO)
(Information to guide
next selection level)

Fig. 15: LVL1 data flow

physics — e.g., they may be produced in Higgs decays. Leptonic W and Z decays give a very clean signature that can be used in the trigger. Of course it is interesting to study W and Z boson production per se, and such events can be very useful for detector studies (e.g., calibration of the EM calorimeters).

In view of the above, LVL1 triggers at hadron colliders search for the following signatures (see Fig. 10).

- High-p_T muons — these can be identified as charged particles that penetrate beyond the calorimeters; a p_T cut is needed to control the rate of muons from $\pi^\pm \rightarrow \mu^\pm\nu$ and $\mathrm{K}^\pm \rightarrow \mu^\pm\nu$ decays in flight, as well as those from semi-muonic beauty and charm decays.
- High-p_T photons — these can be identified as narrow clusters in the EM calorimeter; cuts are made on transverse energy ($E_\mathrm{T} >$ threshold), and isolation and associated hadronic transverse energy ($E_\mathrm{T} <$ threshold), to reduce the rate due to misidentified high-p_T jets.
- High-p_T electrons — identified in a similar way to photons, although some experiments require a matching track as early as LVL1.
- High-p_T taus — identified as narrow clusters in the calorimeters (EM and hadronic energy combined).
- High-p_T jets — identified as wider clusters in the calorimeters (EM and hadronic energy combined); note that one needs to cut at very high p_T to get acceptable rates given that jets are the dominant high-p_T process.
- Large missing E_T or scalar E_T.

Some experiments also search for tracks from displaced secondary vertices at an early stage in the trigger selection.

The trigger selection criteria are typically expressed as a list of conditions that should be satisfied — if any of the conditions is met, a trigger is generated (subject to dead-time requirements, etc.). In these notes, the list of conditions is referred to as the 'trigger menu', although the name varies from experiment to experiment. An illustrative example of a LVL1 trigger menu for high-luminosity running at LHC includes the following (numbers are given for the case of ATLAS):

- one or more muons with $p_\mathrm{T} > 20$ GeV (rate ~ 11 kHz);
- two or more muons each with $p_\mathrm{T} > 6$ GeV (rate ~ 1 kHz);
- one or more e/γ with $E_\mathrm{T} > 30$ GeV (rate ~ 22 kHz);
- two or more e/γ each with $E_\mathrm{T} > 20$ GeV (rate ~ 5 kHz);
- one or more jets with $E_\mathrm{T} > 290$ GeV (rate ~ 200 Hz);
- one or more jets with $E_\mathrm{T} > 100$ GeV and missing-$E_\mathrm{T} > 100$ GeV (rate ~ 500 Hz);

- three or more jets with $E_\mathrm{T} > 130$ GeV (rate ~ 200 Hz);
- four or more jets with $E_\mathrm{T} > 90$ GeV (rate ~ 200 Hz).

The above list represents an extract from a LVL1 trigger menu, indicating some of the most important trigger requirements — the full menu would include many items in addition (typically 100 items in total). The additional items are expected to include the following:

- τ (or isolated single-hadron) candidates;
- combinations of objects of different types (e.g., muon *and* e/γ);
- pre-scaled[5] triggers with lower thresholds;
- triggers needed for technical studies and to aid understanding of the data from the main triggers (e.g., trigger on bunch crossings at random to collect an unbiased data sample).

10 LVL1 trigger design for the LHC

A number of design goals must be kept in mind for the LVL1 triggers at the LHC. It is essential to achieve a very large reduction in the physics rate, otherwise the HLT/DAQ system will be swamped and the dead-time will become unacceptable. In practice, the interaction rate, $\mathcal{O}(10^9)$ Hz, must be reduced to less than 100 kHz in ATLAS and CMS. Complex algorithms are needed to reject the background while keeping the signal events.

Another important constraint is to achieve a short latency — information from all detector elements ($\mathcal{O}(10^7$–$10^8)$ channels!) has to be held on the detector pending the LVL1 decision. The pipeline memories that do this are typically implemented in ASICs (application-specific integrated circuits), and memory size contributes to the cost. Typical LVL1 latency values are a few microseconds (e.g., less than 2.5 μs in ATLAS and less than 3.2 μs in CMS).

A third requirement is to have flexibility to react to changing conditions (e.g., a wide range of luminosities) and — it is hoped — to new physics! The algorithms must be programmable, at least at the level of parameters (thresholds, etc.).

10.1 Case study — ATLAS e/γ trigger

The ATLAS e/γ trigger algorithm can be used to illustrate the techniques used in LVL1 trigger systems at LHC. It is based on 4×4 'overlapping, sliding windows' of trigger towers as illustrated in Fig. 16. Each trigger tower has a lateral extent of 0.1×0.1 in η, ϕ space, where η is pseudorapidity and ϕ is azimuth. There are about 3500 such towers in each of the EM and hadronic calorimeters. Note that each tower participates in calculations for 16 windows. The algorithm requires a local maximum in the EM calorimeter to define the η–ϕ position of the cluster and to avoid double counting of extended clusters (so-called 'declustering'). It can also require that the cluster be isolated, i.e., little energy surrounding the cluster in the EM calorimeter or the hadronic calorimeter.

The implementation of the ATLAS LVL1 calorimeter trigger is sketched in Fig. 17. Analogue electronics on the detector sums signals from individual calorimeter cells to form trigger-tower signals. After transmission to the 'pre-processor' (PPr), which is located in an underground room, the tower signals are received and digitized; then the digital data are processed to obtain estimates of E_T per trigger tower for each BC. At this point in the processing chain (i.e., at the output of the PPr), there is an 'η–ϕ matrix' of the E_T per tower in each of the EM and hadronic calorimeters that gets updated every 25 ns.

[5] Some triggers may be 'pre-scaled' — this means that only every N^{th} event satisfying the relevant criteria is recorded, where N is a parameter called the pre-scale factor; this is useful for collecting samples of high-rate triggers without swamping the T/DAQ system.

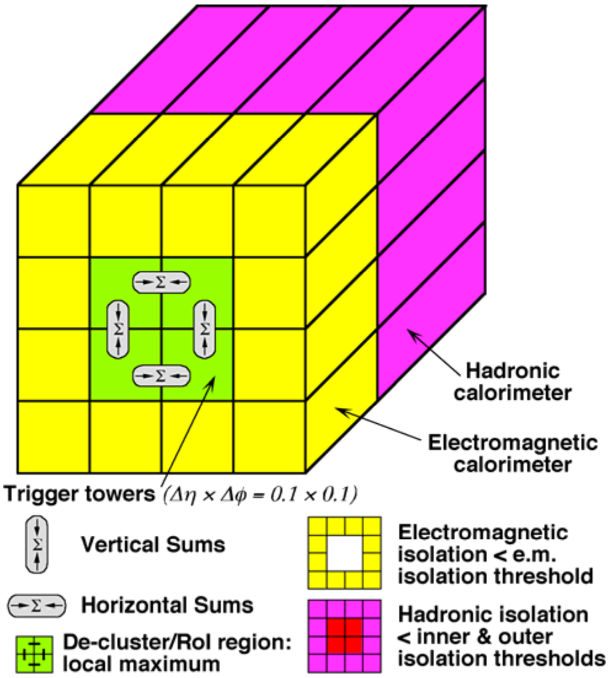

Hadronic calorimeter
Electromagnetic calorimeter
Trigger towers $(\Delta\eta \times \Delta\phi = 0.1 \times 0.1)$
Vertical Sums
Horizontal Sums
De-cluster/RoI region: local maximum
Electromagnetic isolation < e.m. isolation threshold
Hadronic isolation < inner & outer isolation thresholds

Fig. 16: ATLAS e/γ trigger algorithm

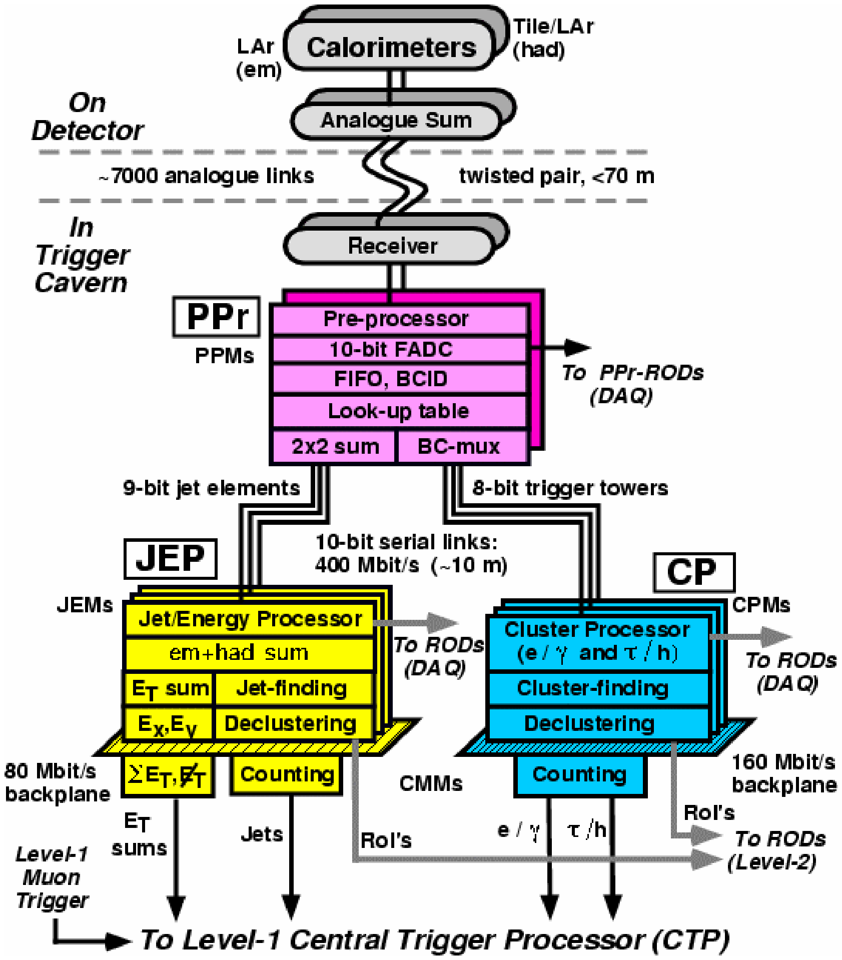


Fig. 17: Overview of the ATLAS LVL1 calorimeter trigger

The tower data from the PPr are transmitted to the cluster processor (CP). Note that the CP is implemented with very dense electronics so that there are only four crates in total. This minimizes the number of towers that need to be transmitted ('fanned out') to more than one crate. Fan out is required for towers that contribute to windows for which the algorithmic processing is implemented in more than one crate. Also, within each CP crate, trigger-tower data need to be fanned out between electronic modules, and then between processing elements within each module. Considerations of connectivity and data-movement drive the design.

In parallel with the CP, a jet/energy processor (JEP) searches for jet candidates and calculates missing-E_T and scalar-E_T sums. This is not described further here.

A very important consideration in designing the LVL1 trigger is the need to identify uniquely the BC that produced the interaction of interest. This is not trivial, especially given that the calorimeter signals extend over many BCs. In order to assign observed energy deposits to a given BC, information has to be combined from a sequence of measurements. Figure 18 illustrates how this is done within the PPr (the logic is repeated ~ 7000 times so that this is done in parallel for all towers). The raw data for a given tower move along a pipeline that is clocked by the 40 MHz BC signal. The multipliers together with the adder tree implement a finite-impulse-response filter whose output is passed to a peak finder (a peak indicates that the energy was deposited in the BC currently being examined) and to a look-up table that converts the peak amplitude to an ET value. Special care is taken to avoid BC misidentification for very large pulses that may get distorted in the analogue electronics, since such signals could correspond to the most interesting events. The functionality shown in Fig. 18 is implemented in ASICs (four channels per ASIC).

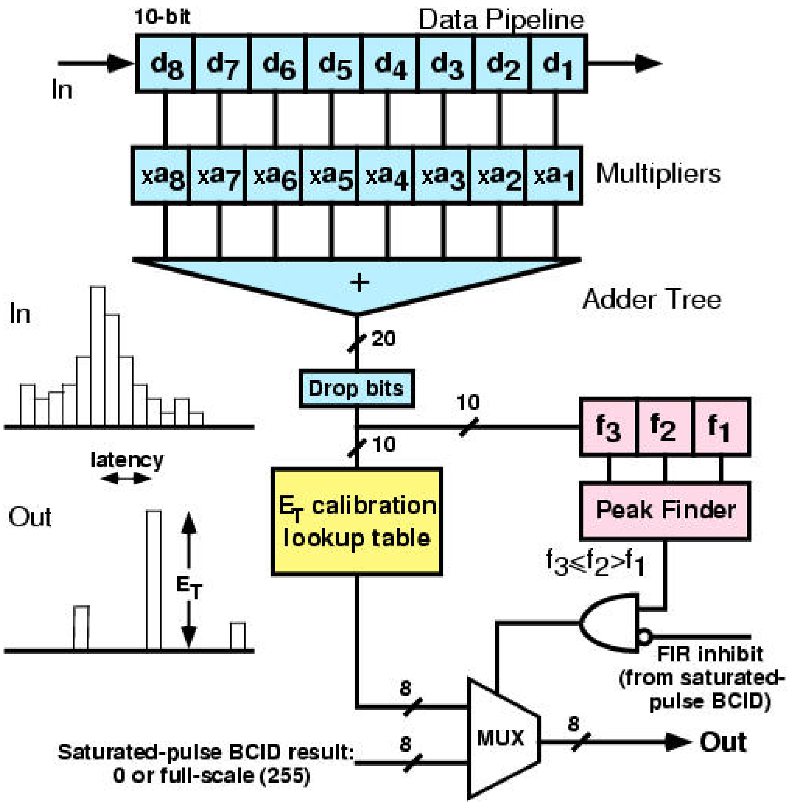


Fig. 18: Bunch-crossing identification

The transmission of the data (i.e., the E_T matrices) from the PPr to the CP is performed using a total of 5000 digital links each operating at 400 Mbits/s (each link carries data from two towers using a technique called BC multiplexing [6]). Where fan out is required, the corresponding links are duplicated with the data being sent to two different CP crates. Within each CP crate, data are shared between neighbouring modules over a very high density crate back-plane (~ 800 pins per slot in a 9U crate; data rate of 160 Mbits/s per signal pin using point-to-point connections). On each of the modules, data are passed to eight large field-programmable gate arrays (FPGAs) that perform the algorithmic processing, fanning out signals to more than one FPGA where required.

As an exercise, it is suggested that students make an order-of-magnitude estimate of the total bandwidth between the PPr and the CP, considering what this corresponds to in terms of an equivalent number of simultaneous telephone calls[6].

The e/γ (together with the τ/h) algorithms are implemented using FPGAs. This has only become feasible thanks to recent advances in FPGA technology since very large and very fast devices are needed. Each FPGA handles an area of 4×2 windows, requiring data from 7×5 towers in each of the EM and hadronic calorimeters. The algorithm is described in a programming language (e.g., VHDL) that can be converted into the FPGA configuration file. This gives flexibility to adapt algorithms in the light of experience — the FPGAs can be reconfigured *in situ*. Note that parameters of the algorithms can be changed easily and quickly, e.g., as the luminosity falls during the course of a coast of the beams in the LHC machine, since they are held in registers inside the FPGAs that can be modified at run time (i.e., there is no need to change the 'program' in the FPGA).

11 High-level triggers and data acquisition at the LHC

In the LHC experiments, data are transferred after a LVL1 trigger accept decision to large buffer memories — in normal operation the subsequent stages should not introduce further dead-time. At this point in the readout chain, the data rates are still massive. An event size of ~ 1 Mbyte (after zero suppression or data compression) at ~ 100 kHz event rate gives a total bandwidth of ~ 100 Gbytes/s (i.e., Ą~ 800 Gbits/s). This is far beyond the capacity of the bus-based event building of LEP. Such high data rates will be dealt with by using network-based event building and by only moving a subset of the data.

Network-based event building is illustrated in Fig. 19 for the example of CMS. Data are stored in the readout systems until they have been transferred to the filter systems [associated with high-level trigger (HLT) processing], or until the event is rejected. Note that no node in the system sees the full data rate — each readout system covers only a part of the detector and each filter system deals with only a fraction of the events.

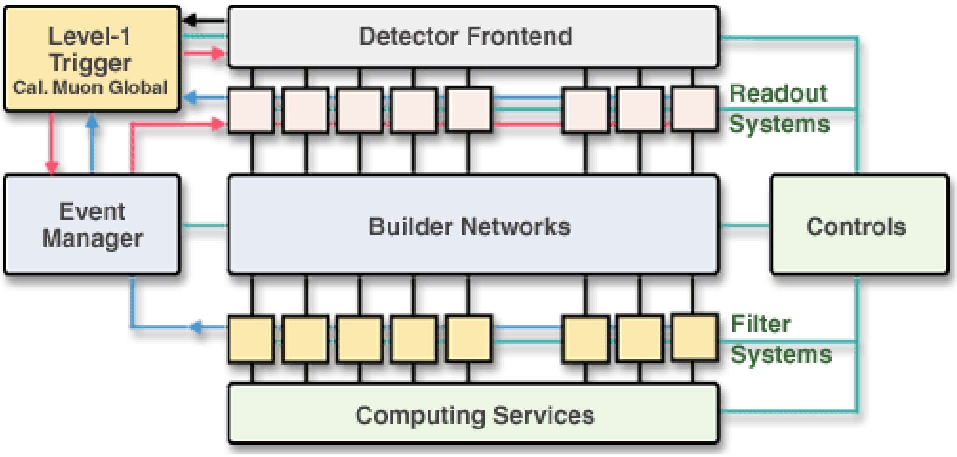


Fig. 19: CMS event builder

[6] You may assume an order-of-magnitude data rate for voice calls of 10 kbits/s — for example, the GSM mobile-phone standard uses a 9600 bit/s digital link to transmit the encoded voice signal.

The LVL2 trigger decision can be made without accessing or processing all of the data. Substantial rejection can be made with respect to LVL1 without accessing the inner-tracking detectors — calorimeter triggers can be refined using the full-precision, full-granularity calorimeter information; muon triggers can be refined using the high-precision readout from the muon detectors. It is therefore only necessary to access the inner-tracking data for the subset of events that pass this initial selection. ATLAS and CMS both use this sequential selection strategy. Nevertheless, the massive data rates pose problems even for network-based event building, and different solutions are being adopted in ATLAS and CMS to address this.

In CMS the event building is factorized into a number of slices, each of which sees only a fraction of the total rate (see Fig. 20). This still requires a large total network bandwidth (which has implications for the cost), but it avoids the need for a very big central network switch. An additional advantage of this approach is that the size of the system can be scaled, starting with a few slices and adding more later (e.g., as additional funding becomes available).

Eight slices:
Each slice sees
only 1/8th of
the events

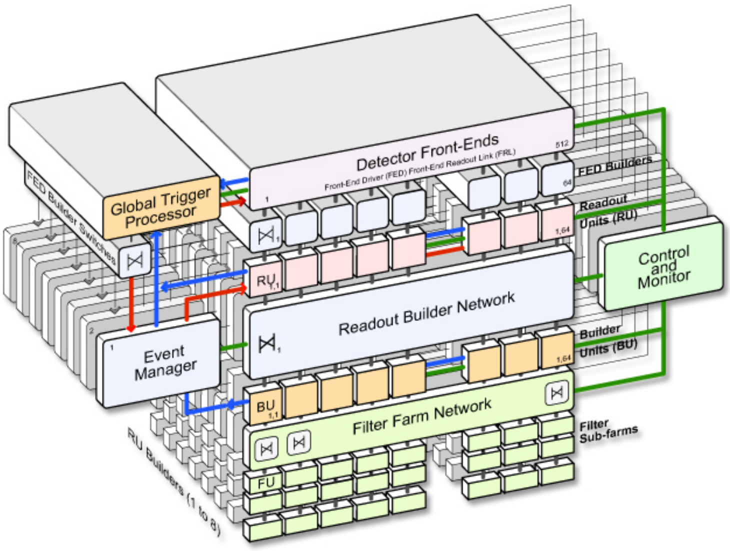


Fig. 20: The CMS slicing concept

In ATLAS the amount of data to be moved is reduced by using the region-of-interest (RoI) mechanism (see Fig. 21). Here, the LVL1 trigger indicates the geographical location in the detector of candidate objects. LVL2 then only needs to access data from the RoIs, a small fraction of the total, even for the calorimeter and muon detectors that participated in the LVL1 selection. This requires relatively complicated mechanisms to serve the data selectively to the LVL2 trigger processors. In the example shown in Fig. 21, two muons are identified by LVL1. It can be seen that only a small fraction of the detector has to be accessed to validate the muons. In a first step only the data from the muon detectors are accessed and processed, and many events will be rejected where the more detailed analysis does not confirm the comparatively crude LVL1 selection (e.g., sharper p_{T} cut). For those events that remain, the inner-tracker data will be accessed within the RoIs, allowing further rejection (e.g., of muons from decays in flight of charged pions and kaons). In a last step, calorimeter information may be accessed within the RoIs to select isolated muons (e.g., to reduce the high rate of events with muons from bottom and charm decays, while retaining those from W and Z decays).

Concerning hardware implementation, the computer industry is putting on the market technologies that can be used to build much of the HLT/DAQ systems at LHC. Computer network products now offer high performance at affordable cost. Personal computers (PCs) provide exceptional value for money in processing power, with high-speed network interfaces as standard items. Nevertheless, custom hardware

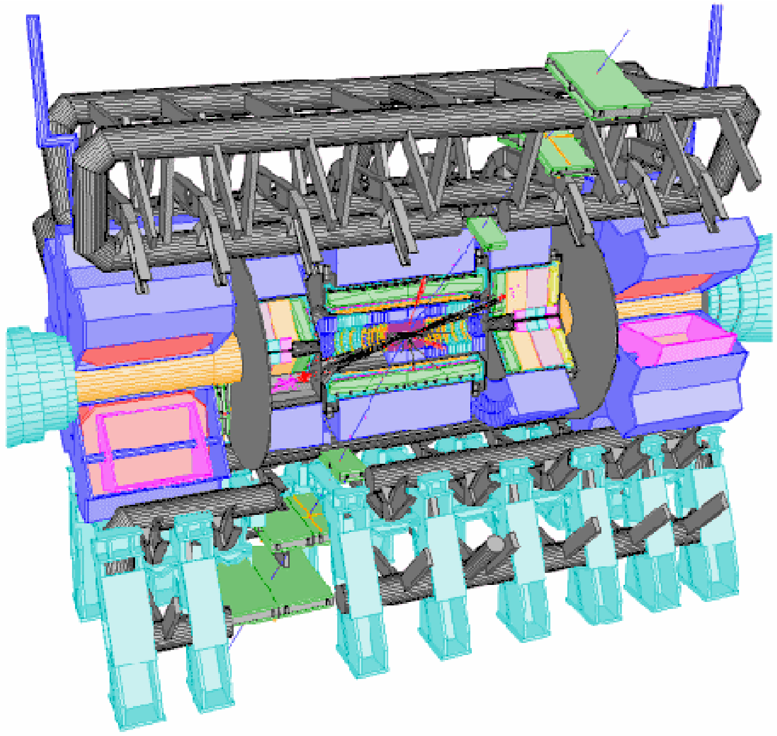


Fig. 21: The ATLAS region-of-interest concept — example of a dimuon event (see text)

is needed in the parts of the system that see the full LVL1 trigger output rate (~ 100 kHz). This concerns the readout systems that receive the detector data following a positive LVL1 trigger decision, and (in ATLAS) the interface to the LVL1 trigger that receives the RoI pointers. Of course, this is in addition to the specialized front-end electronics associated with the detectors that was discussed earlier (digitization, pipelines, derandomizers, etc.).

As for the LVL1 trigger, the HLT has a trigger menu that describes which events should be selected. This is illustrated in Table 2 for the example of CMS, assuming a luminosity for early running of $\mathscr{L} \sim 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The total rate of ~ 100 Hz contains a large fraction of events that are useful for physics analysis. Lower thresholds would be desirable, but the physics coverage has to be balanced against considerations of the offline computing cost. Note that there are large uncertainties on the rate calculations.

Table 2: Estimated high-level trigger rates for $\mathscr{L}$ ~$2 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ (CMS numbers from Ref. [7])

Trigger configuration	**Rate**
One or more electrons with $p_\mathrm{T} > 29$ GeV, or two or more electrons with $p_\mathrm{T} > 17$ GeV	~ 34 Hz
One or more photons with $p_\mathrm{T} > 80$ GeV, or two or more photons with $p_\mathrm{T} > 40, 25$ GeV	~ 9 Hz
One or more muons with $p_\mathrm{T} > 19$ GeV, or two or more muons with $p_\mathrm{T} > 7$ GeV	~ 29 Hz
One or more taus with $p_\mathrm{T} > 86$ GeV, or two or more taus with $p_\mathrm{T} > 59$ GeV	~ 4 Hz
One or more jets with $p_\mathrm{T} > 180$ GeV *and* missing-E_T> 123 GeV	~ 5 Hz
One or more jets with $p_\mathrm{T} > 657$ GeV, or three or more jets with $p_\mathrm{T} > 247$ GeV, or four or more jets with $p_\mathrm{T} > 113$ GeV	~ 9 Hz
Others (electron and jet, b-jets, etc.)	~ 7 Hz

A major challenge lies in the HLT/DAQ software. The event-selection algorithms for the HLT can be subdivided, at least logically, into LVL2 and LVL3 trigger stages. These might be performed by two separate processor systems (e.g., ATLAS), or in two distinct processing steps within the same processor system (e.g., CMS). The algorithms have to be supported by a software framework that manages the flow of data, supervising an event from when it arrives at the HLT/DAQ system until it is either rejected, or accepted and recorded on permanent storage. This includes software for efficient transfer of data to the algorithms. In addition to the above, there is a large amount of associated online software (run control, databases, book-keeping, etc.).

12 Case study — Pierre Auger trigger

The Pierre Auger detectors represent an interesting case for triggering. The detector is distributed over a very large area and a long time would be needed to transmit the signals to a central place. High-speed data links would be expensive; and, of course, there is no cosmic bunch-crossing clock! The basic approach that is used is to have local triggering, independently for each detector unit (tank or fluorescence detector). A reasonably low-rate trigger is needed for each local unit, which must also be efficient for events of interest (i.e., very high energy cosmic rays). When a unit triggers, it sends its local trigger information to the central TDAQ system together with a precise ‘time stamp’. The central system then makes a selection combining all the information that is consistent with a common event time. For selected events, the full data are then collected from all of the relevant detector units. A final level of selection is then made based on a more detailed analysis of the full data.

This approach can be illustrated by considering the Surface Detector (SD). Each tank unit, a photograph of which is shown in Fig. 22, contains three PMTs to detect Cherenkov light produced in the water, local TDAQ electronics, a local control computer, a radio link, and a solar power panel and battery. More details can be found in Ref. [17].

Each SD unit contains a two-level trigger system, data buffering, and a GPS-based time-stamp system (relative precision about 10 ns between units). Signals from the three PMTs in the unit are each digitized by 40 MHz ADCs. (By coincidence this is the same frequency as at the LHC.) The LVL1 trigger analyses the time profile of pulses and requires a coincidence between PMTs, with an output rate below 100 Hz per PMT. It uses programmable logic to implement pipelined trigger processing at 40 MHz.

The LVL2 trigger refines the local selection and achieves an output rate less than 20 Hz per unit using software that runs on the local control processor. The local LVL2 trigger result is sent asynchronously to the central TDAQ over the radio link, while the full data are retained in the local buffer awaiting the LVL3 result.

After receiving trigger information from an SD, the central TDAQ requests trigger information (with looser cuts) from other SD units. The LVL3 trigger combines data from both SD and Fluorescence Detector units. The final event rate depends on the size of the SD array that is active, but is sufficiently low to allow the data to be recorded for offline analysis.

13 Concluding remarks

It is hoped that these lectures have succeeded in giving some insight into the challenges of building T/DAQ systems for HEP experiments. These include challenges connected with the physics (inventing algorithms that are fast, efficient for the physics of interest, and that give a large reduction in rate), and challenges in electronics and computing. It is also hoped that the lectures have demonstrated how the subject has evolved to meet the increasing demands, e.g., of LHC compared to LEP, by using new ideas based on new technologies.

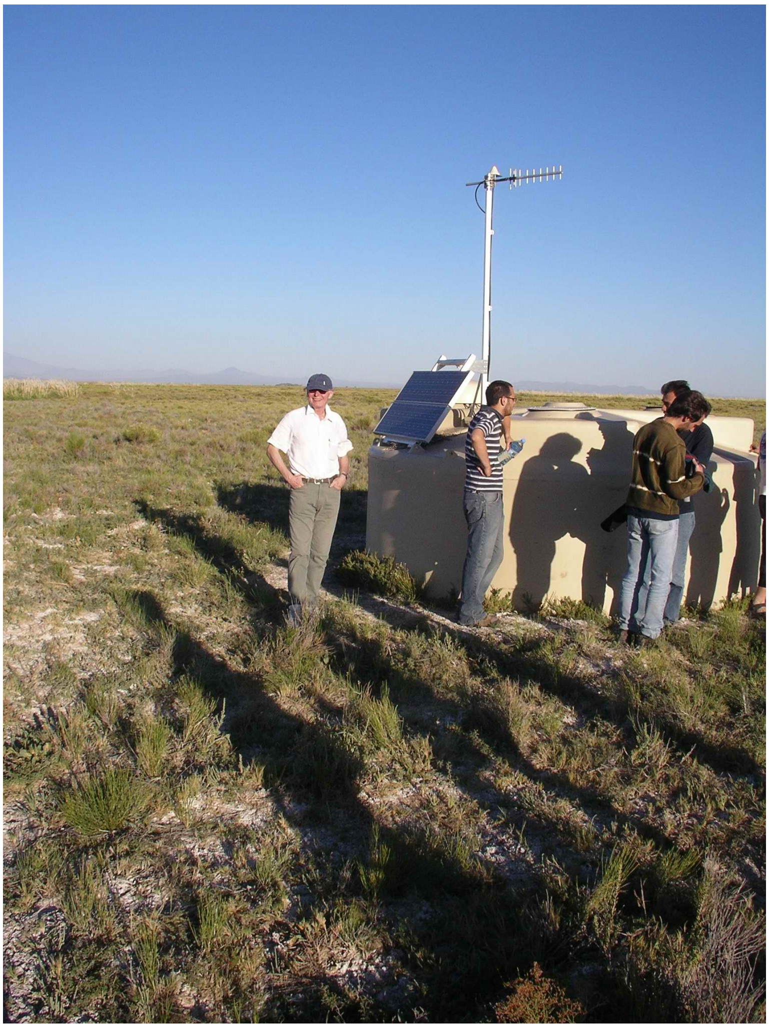


Fig. 22: Photograph of a surface detector unit

Acknowledgements

The author would like to thank the local organizing committee and members of the Pierre Auger Collaboration for their wonderful hospitality during his stay in Argentina. In particular, he would like to thank Tere Dova who, as local director of the school, created such a wonderful atmosphere between all the participants, staff and students alike.

The author would like to thank the following people for their help and advice in preparing the lectures and the present notes: Bob Blair, Helfried Burckhart, Vincenzo Canale, Philippe Charpentier, Eric Eisenhandler, Markus Elsing, Philippe Farthouat, John Harvey, Jim Linnerman, Claudio Luci, Wesley Smith and Alan Watson.

Last, but not least, special thanks are due to Don Gauderio and his friends for all of the fun times we had together during the school.

References

[1] J. Schukraft, Heavy-ion physics at the LHC, in Proc. 2003 CERN-CLAF School of Physics, CERN-2006-001.
ALICE Collaboration, Trigger, Data Acquisition, High Level Trigger, Control System Technical Design Report, CERN-LHCC-2003-062.

[2] W. von Rueden, The ALEPH data acquisition system, *IEEE Trans. Nucl. Sci.*, Vol. 36, No. 5 (1989) 1444–1448.

J. F. Renardy et al., Partitions and trigger supervision in ALEPH, *IEEE Trans. Nucl. Sci.*, Vol. 36, No. 5 (1989) 1464–1468.

A. Belk et al., DAQ software architecture for ALEPH, a large HEP experiment, *IEEE Trans. Nucl. Sci.*, Vol. 36, No. 5 (1989) 1534–1539.

P. Mato et al., The new slow control system for the ALEPH experiment at LEP, *Nucl. Instrum. Methods A* **352** (1994) 247–249.

[3] A. Augustinus et al, The DELPHI trigger system at LEP2 energies, *Nucl. Instrum. Methods A* **515** (2003) 782–799.

DELPHI Collaboration, Internal Notes DELPHI 1999-007 DAS 188 and DELPHI 2000-154 DAS 190 (unpublished).

[4] B. Adeva et al., The construction of the L3 experiment, *Nucl. Instrum. Methods A* **289** (1990) 35–102.

T. Angelov et al., Performances of the central L3 data acquisition system, *Nucl. Instrum. Methods A* **306** (1991) 536–539.

C. Dionisi et al., The third level trigger system of the L3 experiment at LEP, *Nucl. Instrum. Methods A* **336** (1993) 78–90 and references therein.

[5] J.T.M. Baines et al., The data acquisition system of the OPAL detector at LEP, *Nucl. Instrum. Methods A* **325** (1993) 271–293.

[6] ATLAS Collaboration, First-Level Trigger Technical Design Report, CERN-LHCC-98-14.

ATLAS Collaboration, High-Level Triggers, Data Acquisition and Controls Technical Design Report, CERN-LHCC-2003-022.

[7] CMS Collaboration, The Level-1 Trigger Technical Design Report, CERN-LHCC-2000-038.

CMS Collaboration, Data Acquisition and High-Level Trigger Technical Design Report, CERN-LHCC-2002-26.

[8] See, for example, summary talks in Proc. Computing in High Energy and Nuclear Physics, CHEP 2003, `http://www.slac.stanford.edu/econf/C0303241/proceedings.html`

[9] LHCb Collaboration, Online System Technical Design Report, CERN-LHCC-2001-040.

LHCb Collaboration, Trigger System Technical Design Report, CERN-LHCC-2003-031.

[10] S. Stapnes, Instrumentation, in these proceedings.

[11] H1 Collaboration, The H1 detector, *Nucl. Instrum. Methods A* **386** (1997) 310.

[12] R. Carlin et al., The trigger of ZEUS, a flexible system for a high bunch crossing rate collider, *Nucl. Instrum. Methods A* **379** (1996) 542–544.

R. Carlin et al., Experience with the ZEUS trigger system, *Nucl. Phys. B*, Proc. Suppl. **44**(1995) 430–434.

W.H. Smith et al., The ZEUS trigger system, CERN-92-07, pp. 222–225.

[13] CDF IIb Collaboration, The CDF IIb Detector: Technical Design Report, FERMILAB-TM-2198 (2003).

[14] D0 Collaboration, RunIIb Upgrade Technical Design Report, FERMILAB-PUB-02-327-E.

[15] R. Arcidiacono et al., The trigger supervisor of the NA48 experiment at CERN SPS, *Nucl. Instrum. Methods A* **443** (2000) 20–26 and references therein.

[16] T. Fuljahn et al., Concept of the first level trigger for HERA-B, *IEEE Trans. Nucl. Sci.* Vol. 45, No. 4 (1998) 1782–1786.

M. Dam et al., Higher level trigger systems for the HERA-B experiment, *IEEE Trans. Nucl. Sci.* Vol. 45, No. 4 (1998) 1787–1792.

[17] A. Watson, these proceedings.

International Scientific Committee

Carlos García Canal (Universidad Nacional de La Plata, Argentina)
Alvaro De Rújula (CERN, Geneva, Switzerland)
John Ellis (CERN, Geneva, Switzerland)
Nick Ellis (CERN, Geneva, Switzerland)
Robert Fleischer (CERN, Geneva, Switzerland)
Egil Lillestøl (CERN, Geneva, Switzerland), CERN Schools Director
Danielle Métral (CERN, Geneva, Switzerland), School Secretary
Ron Shellar (Centro Brasileiro de Pesquisas Fisicas, Rio de Janeiro, Brazil)
Arnulfo Zepeda (CINVESTAV, Mexico)

Local Organizing Committee

María Teresa Dova (Universidad Nacional de La Plata, Argentina), Local Director
Carlos Kozameh (UNC, Córdoba, Argentina)
Ricardo Piegaia (Universidad de Buenos Aires, Argentina)
Estaban Roulet (Centro Atómico, Bariloche, Argentina)
Óscar Sampayo (Universidad Nacional de Mar del Plata, Argentina)

Lecturers

Luis Álvarez-Gaumé (CERN, Geneva, Switzerland)
John Ellis (CERN, Geneva, Switzerland)
Nick Ellis (CERN, Geneva, Switzerland)
Wick Haxton (University of Washington, Seattle, WA, USA)
Dmitri Kharzeev (BNL, Upton, NY, USA)
Rocky Kolb (Fermilab, Batavia, IL, USA)
Michelangelo Mangano (CERN, Geneva, Switzerland)
Yosef Nir (Weizmann Institute, Rehovot, Israel)
Angela Olinto (University of Chicago, IL, USA)
Paul Sommers (University of Utah, Salt Lake City, UT, USA)
Steinar Stapnes (University of Oslo, Norway and CERN, Geneva, Switzerland)
Alan Watson (University of Leeds, UK)

Discussion Leaders

Guillermo Contreras (CINVESTAV, Merida, Mexico)
Daniel Gómez Dumm (Universidad Nacional de La Plata, Argentina)
Gabriel González Sprinberg (Universidad de la República, Montevideo, Uruguay)
John Swain (Northeastern University, Boston, USA and Universidad de Costa Rica)
Hiroshi Nunokawa (Pontifícia Universidad Católica do Rio de Janeiro, Brazil)

Students

Kazuyoshi Akiba
Leonidas Aliaga Soplín
Rogerio Almeida
Hernán Gonzalo Asorey
Matthew Barrett
José Luis Bazo Alba
Ernesto Alejandro Blanco Covarrubias
Beatriz Blanco Siffert
Alberto Bravo
Nicolas Busca
José Miguel Cabarcas
Óscar Alberto Castillo Felisola
César Manuel Castromonte Flores
Claudio Capa Tira
Jesús Cobos-Martínez
Ismael Cortes Maldonádo
Anne Dabrowski
Caterina Deplano
Sandro Fonseca
Catherine Fry
Carlos García
Félix Francisco González-Canales
Silvia Goy López
Juan Carlos Helo Herrera
Samuel Hevia Zamora
Franciole Marinho
Santiago Andrés Martínez
María Clementina Medina
Walter Mello Jr.
Miguel Mondragón
Allan Morales
Cristina Morales Morales
Juan Cruz Moreno
Alexander Moreno Briceño
Mercedes Paniccia
Laura Paulucci
Eduardo Peinado-Rodríguez
Ricardo Pérez
Carlos Pérez Lara
Michele Pioppi
Juan Racker
Alba Ramírez Rojas
Murilo Rangel
Julia Elizabeth Ruiz Tabasco
Marcial Sánchez Paredes
Matthias Schneebeli
Cristina Schoch Vianna
Fernando Sepúlveda
Emanuele Simili
Luisa Sabrina Stark Schneebeli
Alejandro Tamashiro
Llinersy Uranga
Eric Vázquez-Jauregui

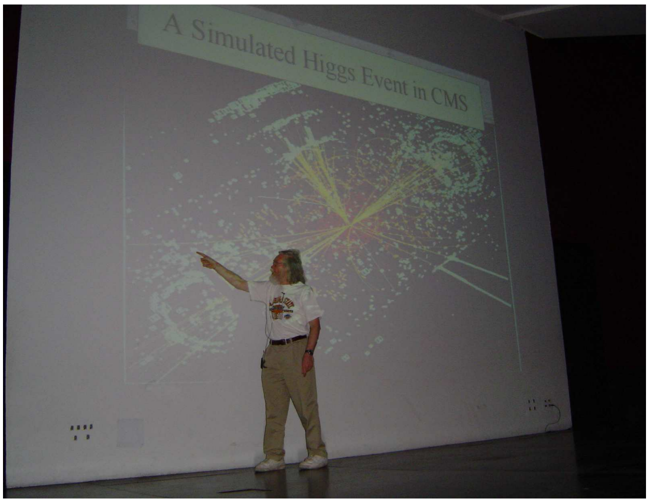
Posters

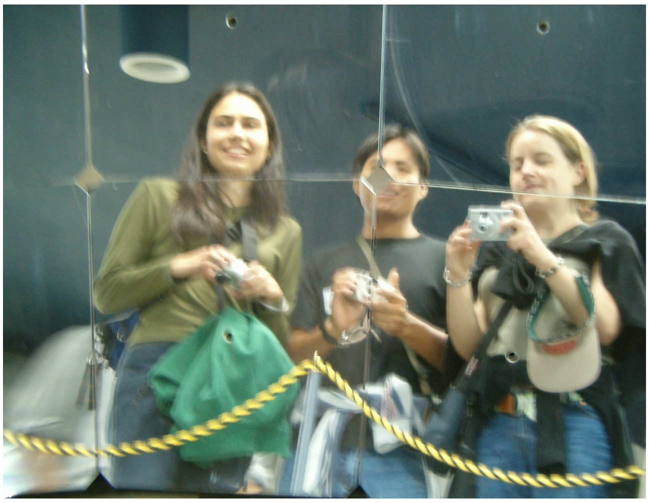
Author	Poster title
Kazuyoshi Akiba	LHCb sensitivity to measure γ (with M. Gandelman)
Leonidas Aliaga Soplín	Testing decoherence effects on neutrino mass measure (with A.M. Gago)
Rogerio Almeida	Study of the Feynman-x and transverse-momentum dependence of Ξ^{*0} and $\overline{\Xi}^{*0}$ production in 230 GeV π^- and K^--nucleon interactions (with F. Rezende, H. da Motta and J.R. Mahón)
Hernán Gonzalo Asorey	Event reconstruction in the surface detector at the UHECR Pierre Auger Observatory (with I. Allekotte)
Matthew Barrett	Dalitz-plot analysis of charmless decays of B-mesons at the BaBar detector
José Luis Bazo Alba	Analysing active galactic nuclei with high-energy neutrinos in IceCube (with A.M. Gago)
Ernesto Alejandro Blanco Covarrubias	Hadroproduction of D^+ and D^- mesons in SELEX (with J. Engelfried)
Beatriz Blanco Siffert	Anisotropy studies in the Pierre Auger Observatory
Nicolas Busca	Analysis of hybrid events for the Pierre Auger Observatory
José Miguel Cabarcas	About the little Higgs model (with R. Martínez)
Óscar Alberto Castillo Felisola	M-theory compactification on G_2 manifolds
César Manuel Castromonte Flores	Study of Λ_c polarization on the $\Lambda_c \to \Lambda\pi$ decay channel (with J.C. Aujos, F.R.A. Simão and E. Polycarpo)
Jesús Cobos-Martínez	MHV amplitudes in QCD
Anne Dabrowski	Measurement of the Cabbibo–Kobayashi–Maskawa quark-mixing matrix element V_{us} from 2003 data from the NA48/2 experiment at CERN (with M. Velasco and M. Szleper)
Caterina Deplano	DIALOG: A front-end chip for the LHCb muon detector (with A. Lai and S. Cadeddu)
Sandro Fonseca	Simulating a forward detector using GEANT4 (with M. Begalli)
Catherine Fry	CALICE-ECAL — Calorimeter R&D for the ILC
Carlos García	Maximal use of kinematic information in the search for the Higgs at D0
Félix Francisco González-Canales	Muon neutrino mass and axial-vector structure in e–ν_μ scattering (with G. Toledo-Sánchez)
Silvia Goy López	The NA48/2 experiment
Franciole Marinho	Analysis of the rare $B^0 \to \mu^+\mu^-$ decay with the reoptimized LHCb detector (with S. Amato and B. de Paula)

Author	**Poster title**
Santiago Andrés Martínez	Synchrotron radiation in space-time foam-based Lorentz-violating electrodynamics (with R. Montemayor, L.F. Urrutia and B. González)
María Clementina Medina	Going down in energy with the Pierre Auger Observatory: it's possible?
Miguel Mondragón	Studies on geometric scaling of the γ^*p cross-section (with G. Contreras)
Allan Morales	Low-energy secondary muons as monitoring instruments of primary cosmic rays (with V. Grabski)
Cristina Morales Morales	Status of $K^{\pm} \to \pi^{\pm}\gamma\gamma$
Juan Cruz Moreno	Studies of surface detector signals in the Pierre Auger Observatory: the rise-time and fall-time dependencies (with M.T. Dova)
Alexander Moreno Briceño	Finite temperature behaviour of an extended model with a triplet Higgs boson (with M. Losada)
Mercedes Paniccia	The Alpha magnetic spectrometer on the international space station
Laura Paulucci	Trapped strangelets in the geomagnetic field (with J.E. Horvath)
Eduardo Peinado-Rodríguez	A minimal S_3-invariant extension of the Standard Model (with A. Mondragón, M. Mondragón, O. Félix, J. Kubo and E. Rodríguez-Jáuregui)
Carlos Pérez Lara	Simulation of the V0 detector for the ALICE project
Michele Pioppi	Tag-side CP violation at BaBar
Murilo Rangel	Studies of exclusive χ_c production with DPEMC
Julia Elizabeth Ruiz Tabasco	Study of scintillating detectors for the V0 detector of ALICE
Marcial Sánchez Paredes	Radiative corrections to the polarized Møller scattering process (with J. Erler)
Matthias Schneebeli	ROME a universally applicable analysis framework generator
Cristina Schoch Vianna	Forward scattering of radio waves by meteors on cosmic-rays — methods and experimental apparatus
Fernando Sepúlveda	Simple electromagnetic cascade simulations (with C. Dib)
Emanuele Simili	Preparation of flow analysis for the ALICE experiment
Luisa Sabrina Stark Schneebeli	Impact of a non-vanishing trilinear scalar coupling A^0 on indirect dark-matter searches (with P. Häfliger)
Alejandro Tamashiro	Effects due to camera shadow on the Auger fluorescence detector absolute calibration (with Auger Drum Calibration Team)
Llinersy Uranga	Theory of point structural particle (with M. Ballester and A. Martínez)
Eric Vázquez-Jáuregui	Hyperon reconstruction study in SELEX (with J. Engelfried)

PHOTOGRAPHS I (MONTAGE)

PHOTOGRAPHS II (MONTAGE)





PHOTOGRAPHS III (MONTAGE)