1 Particle Physics Landscape

In this first chapter, a quick introduction to the Standard Model of particle physics is given. The concepts of elementary particles, interactions and fields are outlined. The experimental side of particle physics is also briefly discussed: how to produce elementary particles, observe them with detectors and make measurements with the data collected by the detectors.

1.1 Elementary Particles

Elementary particle physics, also commonly denoted as high energy physics, is the science that studies the units of matter at the most fundamental level and the nature of the fundamental interactions, the forces, governing their behaviour. Both the units of matter and the interactions are believed to be related to elementary particles. Elementary particles are the simplest objects one can think of: elementary particles have no substructure and thus are not made up of other objects. In the context of the Standard Model of particle physics, elementary particles are supposed to have no spatial extension (confirmed by experiments within the accuracy of their measurements) and are characterised by only a very few quantities: their mass (which could be zero), their spin (the intrinsic angular momentum that could also be zero) and some quantum numbers (like the electric charge) on which the forces depend.

Labelling an object as an elementary particle depends on our ability to probe its possible substructure and thus on our experiments. At the end of the nineteenth century, atomic nuclei were considered elementary. This is no longer the case, as we know that nuclei are made up of protons (discovered by E. Rutherford in the late 1910s) and neutrons (J. Chadwick in 1932), themselves made up of quarks, as revealed by experiments in the 1970s. Modern experiments have not found any substructure of quarks. They are thus considered elementary, a conclusion that could be challenged by future experiments. The electron (e^-), discovered in 1897 by J. J. Thomson, is also considered elementary. Thus, with the electrons and just two species of quarks, named *up* (denoted by the symbol u) and *down* (denoted by the symbol d), all atoms, and hence all ordinary matter observed in nature, can be described. The Standard Model manages to reduce Dmitry Mendeleev's famous periodic table of elements to just three elementary particles!

If elementary particles do not have any constituents, it does not mean that they are themselves necessarily constituents of composite structures. If electrons are the constituents of atoms, and up and down quarks are the constituents of protons and neutrons, respectively

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

(a proton contains two up quarks and one down quark, while a neutron contains two down quarks and one up quark), then other elementary particles do not have the capability of forming larger structures. For instance, the electron neutrino (ν_e) is observed when an unstable nucleus decays by emitting an electron (nuclear beta decay). Neither ν_e nor the electron was present in the nucleus before the decay: their production is the result of the decay itself, allowed by the famous equivalence between energy and mass proposed by Einstein (in modern physics, mass is not a conserved quantity; hence, a heavy particle can potentially decay into lighter ones; only the total energy is conserved). A reason why elementary particles do not necessarily form larger structures is that most of them are very unstable, decaying promptly into lighter elementary particles. For instance, the muon (μ^-) discovered in 1937 is a replica of the electron (it carries the same quantum numbers, spin etc.), except that it is heavier. It decays into lighter particles, while the electron, being the lightest of its species, is necessarily stable. The differences of properties between muons and electrons are then entirely the consequence of the difference in their masses. Another heavy sibling of the electron is the tau (τ^-), discovered in 1975. These three elementary particles strictly carry the same quantum numbers: they all have spin 1/2 in units of $\hbar$, an electric charge $-e$, where e is the charge of the proton, etc. Thus, at a fundamental level, they all have the same interactions;[1] the μ^- and τ^- are simply about 200 and 3 500 times heavier than the electron respectively, leading to slightly different properties. In particle physics, these heavier electrons are said to belong to different *generations*. Each has its own neutrinos: the electron neutrino (ν_e) has already been introduced for the first generation, the muon neutrino (ν_μ) for the second and the tau neutrino (ν_τ) for the third. The three neutrinos share common properties with the e^-, μ^- and τ^- related to the (electro-)weak interaction, as we will discover later. Obviously, they also differ since they are the only fermions of the Standard Model that have no electric charge. The set of these six elementary particles defines the *lepton* family.

Quarks have generations too. As e^- and ν_e, up and down quarks belong to the first generation. Their heavier siblings, the *charm*-quark (c) and *strange*-quark (s), belong to the second generation, and the *top*-quark (t) and *bottom*-quark (b), also called the *truth*-quark and *beauty*-quark, respectively, belong to the third generation. All elementary fermions (spin 1/2 particles) of the Standard Model are listed in Table 1.1 with some of their properties. The six kinds of quarks and leptons are distinguished by *flavour*, i.e. species: there are six flavours of quarks (up, down, strange, charm, bottom and top) and six flavours of leptons (electron, muon, tau, electron neutrino, muon neutrino and tau neutrino).

If the ordinary matter in the universe is made with elementary particles of the first generation, particles from the two other generations have been produced in laboratories, thanks to high energy accelerators. Note that when extreme conditions are encountered in cosmic events, such as a core collapse producing a supernova, those particles must be produced, too. However, since they are highly unstable, only the first generation reaches the earth.[2] It is legitimate to ask whether there are other generations of heavier quarks or leptons. So

[1] Except with the Higgs boson. We will see in Chapter 11 that the interactions of fermions with the Higgs bosons are proportional to their mass.

[2] Muons or composite particles containing a strange quark have been observed from cosmic rays, but they are secondary particles. See Section 1.3.1.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

Table 1.1 Elementary fermions of the Standard Model. Masses are indicated in brackets and expressed in MeV/c^2.

			Generation		
	Spin[a]	Electric charge[b]	1	2	3
Leptons	1/2	0	ν_e (~ 0)[c]	ν_μ (~ 0)[c]	ν_τ (~ 0)[c]
	1/2	-1	e^- (0.511)	μ^- (106)	τ^- (1777)
Quarks	1/2	2/3	u (~ 2)	c ($\sim 1.27 \times 10^3$)	t (172.8×10^3)
	1/2	$-1/3$	d (~ 5)	s (~ 93)	b ($\sim 4.18 \times 10^3$)

[a] In units of $\hbar$. [b] In units of proton charge. [c] See footnote 3.

far, there is no experimental evidence supporting this hypothesis. Moreover, experiments at the Large Electron-Positron collider (LEP) collider concluded in the 1990s that if there is a fourth generation, the mass of the corresponding neutrino must be larger than 45 GeV/c^2 (under some assumptions). Given that the mass of neutrinos of the first three generations is ridiculously tiny[3] compared with other elementary fermions (an upper limit of the order 1 eV/c^2), it seems unlikely to have a fourth generation with such a difference. Therefore, in the Standard Model, only three generations are assumed. At this stage, one can appreciate the similarity between leptons and quarks within a given generation: the difference in electric charge between the two quarks is always equal to one unit (of proton charge), which is also the difference in charge between the two leptons. We shall see that this is a consequence of the symmetric structure of the Standard Model in the following chapters.

One may wonder why we distinguish quarks from leptons. The reason is that both kinds of particles do not experience the same interactions, and thus they present very different properties. Whereas leptons can be observed in their free states (i.e., propagating freely), quarks cannot. Quarks are always confined to bound states that are generically called *hadrons*. Protons or neutrons are two examples of hadrons, but there are hundreds of others. Table 1.2 gives the most common hadrons. One can notice in the table that there are two kinds of hadrons: those containing three quarks form the family of *baryons*, and those containing a quark and an antiquark (antiquarks are denoted by a symbol with a bar over the quark symbol, i.e., $\bar{u}$) called *mesons*. Since this is the first time we encounter an *antiparticle*, it is worth introducing them. A particle (elementary or not) has a corresponding antiparticle with the same mass, lifetime and spin, but opposite internal quantum numbers (electric charge is an internal quantum number, and we shall see that there are others). For instance, the positron (e^+) is the antiparticle of the electron: it is stable, has the same mass as the electron (511 keV/c^2), but a positive electric charge. Some particles are their own antiparticles, for example the π^0 that contains as many quarks as antiquarks of a given flavour (see Table 1.2). The photon (γ) is an example of an elementary particle

[3] In Table 1.1, I oversimplify the notion of neutrino mass, suggesting that ν_e, ν_μ and ν_τ have a definite mass close to zero. Actually, the neutrinos that have this tiny mass are a quantum mechanical superposition of the listed neutrinos, with the latter not having, strictly speaking, a definite mass.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

Table 1.2 Example of hadrons.

	Baryons			Mesons	
Name	Proton	Neutron	Lambda	Pions	Kaons
Symbol	p	n	Λ	π^+, π^0, π^-	$K^+, K^0, \overline{K^0}, K^-$
Quark content	uud	udd	uds	$u\bar{d}$, $\frac{u\bar{u}-d\bar{d}}{\sqrt{2}}$, $\bar{u}d$	$u\bar{s}$, $d\bar{s}$, $\bar{d}s$, $\bar{u}s$
Mass (MeV/c^2)	938	940	1116	140, 135, 140	494, 498, 498, 494

being its own antiparticle. Obviously, only particles having all their internal quantum numbers equal to zero can share this property. When a particle and its antiparticle meet, they annihilate, reduced to pure energy with no residual quantum numbers, from which another pair of particle–antiparticle can emerge. An example is the reaction $e^- + e^+ \rightarrow u + \bar{u}$.

1.2 Fundamental Interactions

1.2.1 Quick Overview

Four fundamental interactions (or forces) are known in nature: gravitational, electromagnetic, weak and strong interactions.

In classical physics, gravitation is the attractive force felt by massive objects and is described by Newton's well-known law. In the context of general relativity, gravity becomes a geometric property of spacetime, which is shaped by the energy and momentum of all possible objects, not only matter but also radiation. As we shall see in this section, gravitation does not have any impact on particles, at least when their energy is far from the Planck scale (defined below). Hence, this book mostly ignores it.

Electromagnetism is the attractive (or repulsive) force felt by objects having opposite (or the same) electric charges and nicely described at the classical level by Maxwell's equations. Its generalisation at the quantum level, the quantum electrodynamics (QED) presented in this book, describes the interaction as an exchange of photons between charged particles. Therefore, all charged particles experience the electromagnetic interactions.

The weak interaction is responsible for many decays of unstable particles. It is the interaction that explains beta radioactivity, where there is the emission of an electron and an anti-neutrino, when a neutron decays into a proton, $n \rightarrow p + e^- + \bar{\nu}_e$. Hence, the weak interaction acts on both quarks and leptons. We shall see that despite their very different manifestations, the electromagnetic and the weak interactions appear as two aspects of a more basic interaction called the electroweak interaction.

Finally, the strong interaction is the force binding quarks in hadrons. Among elementary fermions, only quarks experience this interaction. Its description is quite close in spirit to electromagnetism since in quantum chromodynamics, the theory describing the strong interaction at the quantum level, it arises through an exchange of a massless particle called the gluon.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

Table 1.3 Interactions of elementary fermions.

	Weak	Electromagnetic	Strong
Quarks	□	□	□
$e^{\pm}, \mu^{\pm}, \tau^{\pm}$	□	□	
ν_e, ν_μ, ν_τ	□		

Table 1.3 summarises the interactions to which elementary fermions of the Standard Model are subjected.

Whereas the electromagnetic interaction (and the gravity obviously) is familiar to us in our everyday life, it is not the case of weak and strong interactions. Electromagnetic interaction and gravity are long-range interactions, with the strength of the force decreasing inversely proportional to the distance squared (e.g., in Problem 1.1, Coulomb's law is deduced from Maxwell's equations). Macroscopic objects can then experience these interactions. However, both weak and strong interactions turn out to be only short-range interactions (at the nucleon scale or even less) as we shall see in Sections 1.2.3 and 1.2.4.

In the subatomic world, the gravitational force can be safely ignored compared to the three others for two reasons: first, particles are very light; and second, the Newton constant is also extremely small. For instance, the strength of the gravitational force between two protons, a distance r apart, is

$$f_G = \frac{G_N m_p^2}{r^2} = 5.9 \times 10^{-39} \frac{\hbar c}{r^2},$$

with $G_N \sim 6.7 \times 10^{-39} \hbar c \ (\mathrm{GeV}/c^2)^{-2}$. On the other hand, the strength of the electromagnetic force between those protons is

$$f_{\mathrm{EM}} = \frac{e^2}{4\pi\epsilon_0 r^2} = \alpha \frac{\hbar c}{r^2} = \frac{1}{137} \frac{\hbar c}{r^2},$$

with $\alpha \sim 1/137$. (See Table 1, p. xix for the expressions and numerical values of the constants). The ratio of the two shows that the gravitational force is about 10^{36} times weaker than the electromagnetic force! Notice that both interactions would be of the same order of magnitude, i.e. $\hbar c/r^2$, for masses about the Planck mass $m_{Pl} = \sqrt{\hbar c/G_N} \sim 1.2 \times 10^{19}\ \mathrm{GeV}/c^2$. Such masses or equivalently such energies[4] are far beyond the reach of any modern accelerator or known cosmic events (the cosmic ray with the highest energy ever seen on earth is about 3×10^{11} GeV and the most powerful accelerator barely boosts protons to 7×10^3 GeV). If the electromagnetic force is so much stronger than gravity, then it might be surprising that, in everyday life, the effects of gravity seem to be rather dominant. The reason for this is simply that in electromagnetism, positive and negative electric charges compensate each other, with the matter being globally neutral, whereas the gravitational force increases with mass without any such compensation.

There is another more worrying reason why the gravitational force does not belong to the corpus of the Standard Model of particle physics. As we will see in the next three sections,

[4] In general relativity, the gravity's source is the energy–momentum tensor, affecting the spacetime curvature, not only the mass, as in Newtonian physics.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

interactions are described in terms of the exchange of elementary particles using quantum field theory as a theoretical framework. Unfortunately, there is not yet a well-established quantum theory of gravitation! Candidates for such a theory to reconcile quantum mechanics and general relativity, such as string theory or loop quantum gravity, are still unproven. Consequently, only the electromagnetic, weak and strong interactions are described by the Standard Model of particle physics.

1.2.2 Need for Fields

Modern physics introduces the notion of fields to avoid the issues of forces capable of acting at a distance instantly. As an example, imagine that suddenly the sun vanishes. Since in modern physics, nothing can propagate faster than light, the effect of the sun disappearance must take at least eight minutes before reaching the earth, and thus the gravitational force due to the sun must continue to act during this interval even if the sun is no longer present. Hence, the sun itself is not enough to explain the transmission of the gravitational force. This role is played by the gravitational field. Similarly, there is an electromagnetic field, a strong field and a weak field associated with these interactions. These fields permeate all of space. They respond locally to sources (masses for classical gravitation, electric charges for electromagnetism, etc.) and act on another distant point x, propagating the action to it at a finite velocity (at most the speed of light). The force felt at x thus results from the state of the field *locally* at that point. The action at a distance of classical physics is then avoided.

If fields are the natural consequence of the principle of relativity in modern physics, modern physics also includes another key ingredient, quantum mechanics, which implies that fields can only exist in well-defined states of definite quantised energy. This is very similar to the harmonic oscillator quantisation with which the reader should already be familiar. The next conceptual step is to identify those quanta as particles. At the beginning of the twentieth century, Einstein was the first to realise that the photon was the quantum of the electromagnetic field, manifesting a particle-like aspect in the photoelectric effect. It took about 80 more years to identify the quanta of the weak and strong interactions – the $W^{\pm}$ and Z^0 bosons for the weak interaction and the gluons for the strong interaction. These bosons (they are spin 1 particles) are the force carriers of the interactions, with the action of a force being the result of the exchange of quanta (i.e., the force-carrier particles) of the associated field. The following chapters will be mostly devoted to their presentation: quantum numbers carried by the fields, equation of propagation of the field, etc. Table 1.4 summarises some of their properties.

Notice that because of quantum mechanics, even elementary fermions have a wave-like or field aspect. Therefore, the classical conceptual distinction between matter and forces based on the notion of particles on the one side and fields on the other is no longer relevant. All particles, elementary fermions or force carriers, must be described in terms of quanta of fields. The appropriate theoretical framework is then the relativistic quantum field theory, briefly introduced in Chapters 5 and 6. It turns out that the force fields (and their quanta) emerge naturally from the necessity to respect some particular symmetries at the *local* level: physics must be invariant under such transformations that depend on the spacetime point (as opposed to a *global* symmetry transformation that acts on every spacetime point

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

Table 1.4 Elementary bosons of the Standard Model.

		Spin[a]	Electric charge[b]	Multiplicity	Mass (GeV/c^2)	Force carrier of
Photon	γ	1	0	1	0	Electromagnetism (QED)
W bosons	$W^{\pm}$	1	± 1	2	80.4	Weak force
Z boson	Z^0	1	0	1	91.2	Weak force
Gluons	g	1	0	8	0	Strong force (QCD)
Higgs	H	0	0	1	125.2	None

[a]In units of $\hbar$. [b]In units of proton charge.

in the same manner). Then, the quanta (the force carriers) of the force fields described by the Standard Model of particle physics must be spin 1 particles (in units of $\hbar$). For instance, anticipating Chapter 6, we will see that the gauge invariance of Maxwell's equations, probably already known by the reader, is connected to the local phase invariance of the fermion fields. Indeed, the local phase invariance imposes a specific transformation of the electromagnetic field (whose quantum is the photon, a spin 1 boson) that precisely leaves invariant Maxwell's equations. This is an example of a gauge theory that will be further generalised to more complicated symmetries for the other interactions (Chapters 8 and 10 of this book).

1.2.3 Yukawa's Theory of Short-Range Interactions

Historically, the strong force was first discovered as the nuclear force binding protons and neutrons in the atomic nucleus. A model due to the Japanese physicist Hideki Yukawa (1935) assumed a potential of the form

$$\varphi(\boldsymbol{r}) = \frac{g}{4\pi}\frac{e^{-r/r_0}}{r}, \tag{1.1}$$

where $r = |\boldsymbol{r}|$, r_0 is a representative parameter of the range of the interaction, and g is a constant analogous to the electric charge, representative of the strength of the interaction. For $r > r_0$, the potential becomes rapidly negligible, leading to an interaction range of the order of r_0. Since the range of the nuclear force is at most the size of a few protons (the nuclear force is responsible for the cohesion of nuclei but not for the cohesion of larger structures), r_0 must be of the order of a few fm. Let us emphasise the difference between Eq. (1.1) and the familiar electrostatic scalar potential due to a point-like charge, e. Since in electrostatics, the electric field simply satisfies $\boldsymbol{E} = -\nabla V$, the Maxwell equation $\nabla \cdot \boldsymbol{E} = \rho/\epsilon_0$ with $\rho(\boldsymbol{r}) = e\delta(\boldsymbol{r})$ implies Poisson's equation

$$\nabla^2 V = -\frac{e}{\epsilon_0}\delta(\boldsymbol{r}),$$

whose solution (Problem 1.2) is the electrostatic potential

$$V(\boldsymbol{r}) = \frac{e}{4\pi\epsilon_0}\frac{1}{r}. \tag{1.2}$$

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

It depends only on $1/r$ and thus allows a long-range interaction. In addition, unlike $V(\boldsymbol{r})$, $\varphi(\boldsymbol{r})$ in Eq. (1.1) is a solution (Problem 1.2) of

$$\left(\nabla^2 - \frac{1}{r_0^2}\right)\varphi = -g\delta(\boldsymbol{r}).$$

The parameter r_0 is naturally related to the characteristic mass

$$m_0 = \frac{\hbar}{r_0 c} \tag{1.3}$$

(check the consistency of the units), so that

$$\left(\nabla^2 - \left(\frac{m_0 c^2}{\hbar c}\right)^2\right)\varphi = -g\delta(\boldsymbol{r}).$$

Yukawa then extended this equation to the non-static case and interpreted $\varphi(\boldsymbol{r}, t)$ as a quantised field satisfying in the vacuum

$$\left(\nabla^2 - \frac{\partial^2}{c^2 \partial t^2} - \left(\frac{m_0 c^2}{\hbar c}\right)^2\right)\varphi(\boldsymbol{r}, t) = 0. \tag{1.4}$$

The parameter m_0 is then the mass associated with the quanta of the field $\varphi(\boldsymbol{r}, t)$. The reader might recognise the relativistic Klein–Gordon equation; more details about this equation and its quantisation are given in Chapters 5 and 6. Hence, in this model, a short-range interaction implies a massive field and, thus, a massive carrier of the interaction. In his publication (Yukawa, 1935), Yukawa set the value $r_0 = 2$ fm and hence predicted, according to Eq. (1.3), that the quantum of the field φ is a new particle with a mass $m_\varphi = m_0 = 100\ \mathrm{MeV}/c^2$. He observed that the scattering of a neutron by a proton $n + p \to n + p$ could be described by the combination of two elementary processes: the incident neutron emits the quantum particle φ and becomes a proton, $n \to p + \varphi$, while the incident proton absorbs the quantum-particle and becomes a neutron $p + \varphi \to n$. He stated that the exchanged quantum particle 'cannot be emitted into the outer space' since[5] $m_n \ll m_p + m_\varphi$. In modern language, in this scattering, Yukawa stated that the intermediate particle carrying the interaction must be a *virtual particle*. Virtuality is allowed by the uncertainty principle of quantum mechanics

$$\Delta E \Delta t \gtrsim \hbar,$$

whose interpretation is not straightforward [see, e.g., (Aharonov and Bohm, 1961) for an interesting discussion]. Here, if the intermediate state $p + \varphi$ has a lifetime Δt (the time interval between the emission of φ and its absorption), its energy uncertainty is of the order of ΔE given by the above relation. For two nucleons separated by a distance r, and assuming a velocity of φ of the order of c (which is, of course, an upper bound), $\Delta t \sim r/c$,

[5] In the rest frame of φ, if the proton has a momentum $\boldsymbol{p}$, the neutron has $\boldsymbol{p}$ and denoting by p_n and p_p the 4-momentum of the neutron and the proton, respectively, the conservation of the energy–momentum requires

$$m_\varphi^2 = (p_n - p_p)^2 = m_n^2 + m_p^2 - 2\left(\sqrt{m_n^2 + |\boldsymbol{p}|^2}\sqrt{m_p^2 + |\boldsymbol{p}|^2} + |\boldsymbol{p}|^2\right) \leq (m_n - m_p)^2,$$

which is not possible for $m_\varphi = 100\ \mathrm{MeV}/c^2$, given that $m_n \sim m_p \sim 940\ \mathrm{MeV}/c^2$.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

and thus $\Delta E \gtrsim \hbar c/r$. When r is small enough, ΔE can reach $m_\varphi c^2$ without contradicting the energy conservation principle. This implies that r is at most $\hbar c/m_\varphi c^2 = r_0$. The parameter r_0 can then be seen as the range of the interaction. While the presence of the quantum particle φ cannot be observed directly in this scattering, Yukawa's theory implies its existence, and some kinematically possible processes should allow its production and detection. Two years after Yukawa's paper, while analysing cosmic rays, Anderson and Neddermeyer (1937, 1938), followed shortly by Street and Stevenson (1937), discovered a new charged particle whose mass was around 100 MeV/c^2. It was a perfect candidate for φ. However, a few years later, experiments showed that these cosmic ray particles barely interacted with nuclei, contrary to expectation. It turns out that this new particle was actually the muon. It was only in 1947 that, while analysing cosmic rays, Lattes et al. (1947) found a heavier particle, with a mass of about 140 MeV/c^2, interacting with nuclei in a manner consistent with Yukawa's prediction. This particle is known today as the pion, with its three charge states denoted as π^+, π^- and π^0.

1.2.4 A Glimpse of the Weak and Strong Interactions

Yukawa's theory relates the range of interactions with the mass of the force carrier via Eq. (1.3). With masses of the order of 100 GeV/c^2 (see Table 1.4), the carriers of the weak interaction, $W^\pm$ and Z^0, imply an interaction range of about 10^{-18} m. The pions, being 1 000 times lighter, imply a nuclear interaction range of about 10^{-15} m. They can be considered as the carriers of the effective strong force at the scale of the nucleon. Notice that pions are spin 0 bosons and not elementary particles since they are mesons. Therefore, they cannot be the carriers of the strong interaction at the most fundamental level, i.e. at the scale of the quarks. Those are the gluons, and as expected they are spin 1 elementary particles. However, the attentive reader will have remarked that they are massless (cf. Table 1.4). When the force carrier of an interaction is a massless particle, according to Eq. (1.3), we would expect an infinite range. The electromagnetic interaction is clearly active over a very large range since the light of distant galaxies can be observed on earth. However, massless gluons seem to contradict the apparent short-range behaviour of the strong interaction. In order to resolve this contradiction, one should dive deep into the theoretical framework of the strong interaction, quantum chromo dynamics (QCD), a theory presented in Chapter 8. A simplified summary is given here.

When two quarks strongly interact, they exchange gluons as they similarly exchange photons in electromagnetic interactions (QED). The role played by the electric charge in QED is replaced by another kind of charge, the *colour*. If there is only one elementary electric charge e (all charges of particles are just a – possibly fractional – multiple of e), quarks can have three 'colour charges', arbitrarily called *red*, *green* and *blue*. Hence, a quark is labelled by its flavour but also its colour, for instance, u_r for a red up quark. Since the gluons are massless particles, there is intrinsically no theoretical limit to the range of the interaction. However, there is a specific property of QCD, the *colour confinement*, that prevents the observation of quarks as free isolated particles. Experimentally, even by smashing two hadrons at very high energy, detectors observe hadrons again. The higher the energy, the greater the number of hadrons that is detected, without producing isolated quarks. QCD

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

asserts that all particle states observed in nature must be 'colourless' (the appropriate term is colour singlet coming from group theory, meaning that a rotation in the colour space leaves the state invariant, but at this stage of the book, colourless is probably more suitable). Of course, a single quark carrying a colour cannot be colourless. But mesons made of a quark and an antiquark can: if the quark has a colour, the antiquark has an anti-colour. For example, if the quark is red, it is sufficient that the antiquark is anti-red to cancel the overall colour of the quark–antiquark system. Baryons are also colourless because their three quarks have three different colours: inspired by the colour theory of visual arts,[6] the superposition of red, green and blue produces the achromatic white. (From this perspective, an anti-colour can be seen as the complementary colour.) How can the colour confinement be qualitatively explained, and why does it have an impact on the range of the strong interaction? We saw that the exchange of a massless boson leads naturally to a Coulomb-like potential with a $1/r$ dependence, incompatible with confinement. At very short distances ($< 10^{-15}$ m), the QCD colour potential between a quark q_1 and an antiquark $\bar{q}_2$, separated by the distance r, is indeed in $1/r$, adequately described by the exchange of a gluon. However, for longer distances, there is a linear behaviour, leading to the phenomenological potential

$$V(r) = -\frac{A}{r} + \sigma r, \tag{1.5}$$

where A and σ are two positive constants. Therefore, the stored potential energy between the $q_1\bar{q}_2$ pair increases with distance and can become so high that it exceeds the energy required to produce a new pair of $q_3\bar{q}_4$. It costs less energy to form the two pairs $q_1\bar{q}_4$ and $q_3\bar{q}_2$ because the distance between the quark and the antiquark of these pairs is reduced with respect to the initial distance between q_1 and $\bar{q}_2$. Hence, quark–antiquark systems are necessarily confined. This mechanism is, however, not yet quantitatively understood in detail. At a larger scale, the residual nuclear force between hadrons can be understood as follows: far from a hadron, the colour of the quarks constituent of the hadron is not seen, since the quark configuration is such that the hadron is colourless. Therefore, the strong interaction between hadrons rapidly vanishes, producing a short-range nuclear force. In Chapter 8, when we are more familiar with quantum chromodynamics and its Feynman diagrams, we will see how we can recover the pion model exchange from gluon emissions. It is often said that nuclear force is analogous to the Van der Waals forces, where electric dipoles provide intermolecular bondings. Actually, if a similar effect could exist with colour charge instead of electric charge, its magnitude would be too weak to explain the bonding between nucleons (Povh et al., 2008, Chapter 16). The mediation via pions (and possibly heavier mesons) is really required to describe the nuclear force.

After its range, another aspect characterising an interaction is its strength. If it was easy to define the strength of the gravitation and electromagnetic interactions in Section 1.2.1, it is less obvious for the weak and strong interactions. The weak interaction is called 'weak' because its strength appears to be the weakest of the three interactions described by the Standard Model. However, we shall see in Chapter 9 that intrinsically, the interaction is

[6] The visual arts provide a simple analogy for colour combination, but the exact rules are governed by group theory.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

Table 1.5 Decays of Λ and Σ baryons.

Baryons	Quark contents	Mass (GeV/c^2)	Lifetime (s)	Decay channels
Λ^0	*uds*	1 116	2.6×10^{-10}	$p + \pi^-, n + \pi^0$
Σ^0	*uds*	1 193	7.4×10^{-20} † #	$\Lambda^0 + \gamma$
$\Sigma^0(1\,385)$	*uds*	1 384	$\sim 2 \times 10^{-23}$ † ♭	$\Lambda^0 + \pi^0$, $\Sigma^0 + \pi^0$

† Such a short lifetime cannot be measured from the decay length distribution in detectors.

The Σ^0 lifetime can be indirectly measured via the cross-section $\Lambda + \gamma \rightarrow \Sigma^0 + \gamma$, where the γs come from the Coulomb field of a nucleus (Dydak et al., 1977).

♭ The $\Sigma^0(1\,385)$ lifetime can only be estimated from the intrinsic energy uncertainty of unstable states, via $\tau \sim \hbar/\Delta E$ (see Section 3.4).

not that weak, in particular at high energy. Similarly, the strength of the strong interaction (the strongest of the three) depends on the energy scale: we will see in Chapter 8 that at high energy, i.e. at very small distances, its strength is rather weak, while for distances of the order of the size of the hadrons, its strength is extremely powerful. However, even if the notion of interaction strength becomes complicated, a crude estimate can be inferred at low energy by comparing the lifetimes of similar particles that decay mostly in a decay channel involving a specific interaction. Intuitively, the stronger the interaction, the higher the probability of decay, and therefore the shorter the particle's lifetime. Let us introduce two dimensionless constants, α_w and α_s, for the weak and strong interactions, respectively. They are analogous to α, the fine structure constant of electromagnetism. We will learn how to calculate the lifetimes of particles in the following chapters, but at this stage, let us admit that if the boson, carrier of the interaction, is emitted or absorbed, the lifetime formula gets a factor proportional to the inverse of the constant. In Table 1.5, three baryons containing the same quark content (*uds*) are compared. They differ by their quantum numbers. Both Λ^0 and Σ^0 are spin 1/2 particles, while $\Sigma^0(1\,385)$ has a spin 3/2. The difference between Λ^0 and Σ^0 is the isospin quantum number[7] since $I = 0$ for Λ^0 and 1 for Σ^0. Their lifetime and their most frequent decay channels, accounting for almost 100% of the cases, are listed in the table. The only interaction allowing a change of generation among the elementary fermions is the weak interaction. The decays $\Lambda^0 \rightarrow p + \pi^-$ and $\Lambda^0 \rightarrow n + \pi^0$ proceed then via this interaction. At the quark level, the strange quark of the Λ^0 decays as $s \rightarrow u + W^-$, followed by $W^- \rightarrow d + \bar{u}$. The virtual W^-, carrier of the weak interaction, is then emitted and absorbed, yielding a factor α_w^{-2} in the Λ^0 lifetime formula. On the other hand, the decay $\Sigma^0 \rightarrow \Lambda^0 + \gamma$ involves the electromagnetic interaction since a real photon is emitted, but not absorbed. It contributes to a factor α^{-1} to the Σ^0 lifetime. Hence, the ratio of those two lifetimes satisfies

$$\frac{\tau_{\Lambda^0}}{\tau_{\Sigma^0}} \propto \frac{\alpha}{\alpha_w^2}.$$

[7] Isospin symmetry (introduced in Section 8.1.1) assumes that the up and down quarks belong to a two-component vector of isospin value $I = 1/2$. The up quark is the upper component with $I_z = +1/2$, and the down quark is the lower component with $I_z = -1/2$. The maths of the isospin is the same as that of the usual spin. The up and down quarks were so named because of that symmetry.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

In order to get a crude estimate of the constants, we will assume that the previous formula is actually an equality, neglecting all other differences in the lifetime formula (like the phase space). We conclude, using the value $\alpha \sim 1/137$, that

$$\alpha_w \sim \sqrt{\alpha \frac{\tau_{\Sigma^0}}{\tau_{\Lambda^0}}} \sim 10^{-6}.$$

The weak interaction is thus about 10^4 weaker than the electromagnetism interaction. Finally, the decay of $\Sigma^0(1\,385)$ is a strong decay that conserves the isospin numbers (the π^0 has $I = 1$ and $I_z = 0$). One of the quarks of the $\Sigma^0(1\,385)$ emits a virtual gluon that converts into a pair $u\bar{u}$ or $d\bar{d}$, becoming the π^0. Hence, we expect $\tau_{\Sigma^0(1\,385)} \propto \alpha_s^{-2}$, allowing us to conclude

$$\alpha_s \sim \sqrt{\alpha \frac{\tau_{\Sigma^0}}{\tau_{\Sigma^0(1\,385)}}} \sim 1\text{–}10.$$

In summary, at low energy, the strength of the interactions of the Standard Model is roughly

$$\begin{array}{ll} \text{strong:} & \alpha_s \sim 1\text{–}10, \\ \text{electromagnetism:} & \alpha = 1/137 \sim 10^{-2}, \\ \text{weak:} & \alpha_w \sim 10^{-6}. \end{array} \tag{1.6}$$

Do not take those numbers too seriously: other decay channels may lead to different estimates. They are just indicative of the order of magnitude of the interaction strengths. We can, however, keep in mind that the lifetime of particles whose decay is dominated by a single interaction is typically

$$\begin{array}{ll} \text{strong:} & \tau_s \sim 10^{-24}\text{–}10^{-21}\ \text{s}, \\ \text{electromagnetism:} & \tau_{\text{E.M.}} \sim 10^{-20}\text{–}10^{-16}\ \text{s}, \\ \text{weak:} & \tau_w \sim 10^{-13}\text{–}10^{3}\ \text{s}. \end{array} \tag{1.7}$$

1.3 Production of Particles

To study particles, the first step is to produce them. Two mechanisms can be identified: natural production originating from cosmic rays or artificial production requiring an accelerator.

1.3.1 Cosmic Rays

Until the early 1950s, the only source of high energy particles was the interaction of primary cosmic rays with the atmosphere. Primary cosmic rays are charged particles, whose lifetime is long enough to be produced by astrophysical sources and reach the earth. They are mostly protons, and nuclei, whose relative abondance depends on the energy range considered. For example, at about 10 GeV per nucleon, 94% are protons, 6% helium nuclei (alpha particle) and the remaining 1% are stable nuclei synthesised in stars (Particle Data Group, 2022).

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

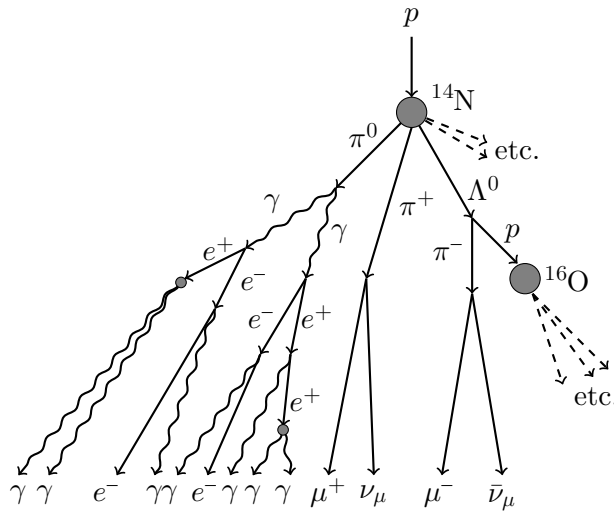


Fig. 1.1 Development of cosmic rays air shower. The small grey circles are electrons in the atmosphere.

By entering the stratosphere, primary cosmic rays cause nuclear collisions with mostly oxygen and nitrogen atoms. These collisions produce secondary particles, including protons, neutrons and charged mesons, such as pions ($\pi^{\pm}$) and kaons ($K^{\pm}$), which typically decay into muons and neutrinos. Since the lifetimes of charged pions and kaons are large enough to travel over significant distances, some may collide with another nucleus in the air before decaying. Neutral pions (π^0) are also produced from the primary interaction. They typically decay almost immediately into a pair of high energy photons that interact electromagnetically with the Coulomb field of atoms. This results in pairs of electron and positron, which can further radiate photons (gamma rays). This process continues, creating a cascade of electrons, positrons and photons. Figure 1.1 illustrates the interaction of a primary cosmic rays in the atmosphere. The number of particles increases rapidly as the cascade of particles, also called *shower*, moves downwards. In each interaction, the particles lose energy, and eventually will not be able to create new particles. After some point, more particles are stopped than created, and the number of shower particles declines.

Energetic secondary particles reaching sea level are dominated by neutrinos and muons. To give an order of magnitude, the neutrino flux around 1 GeV (where the probability of interaction is maximum) is about 1 $\text{cm}^{-2}\ \text{s}^{-1}$ from all directions (Gaisser, 1990). For muons with mean energy above 1 GeV, the intensity of the vertical flux is $I_v \simeq 7 \times 10^{-3}\ \text{cm}^{-2}\ \text{s}^{-1}\ \text{sr}^{-1}$ (Grieder, 2001). Given that the muon flux intensity varies with the zenith angle, θ, as $I(\theta) \simeq I_v \cos^2\theta$ (empirical relationship), the muon flux coming from the sky collected by a horizontal detector is then of the order of 1 $\text{cm}^{-2}\ \text{mn}^{-1}$ (Problem 1.3).

Before the 1950s, it was the analysis of the secondary particles recorded in cloud chambers at the ground level or detected in photographic emulsions flown in a balloon in the upper atmosphere that led to the discoveries of the positron (1932), muon (1937), pion (1947), kaon (1947) and Λ^0 (1950).

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

1.3.2 Accelerators

After the 1950s came the era of particle accelerators. They have the great advantage to provide controlled collisions: the energy of the incident particle and the projectile is reasonably determined, as well as their species, and high energy collisions can be reliably repeated. The technology of accelerators is complex and deserves a book on its own. See, for example, Wiedemann (2007). I limit myself here to very general considerations.

Cyclotrons were first built: they consist of two D-shaped cavities, called dees, facing each other (ⅅ) in which a magnetic field bends the trajectory of a charged particle on a circular path. The dees are kept at different electrical potentials to accelerate the particle as it passes from one dee to the other. The process is repeated twice per turn and therefore requires swapping the potentials of the two dees to maintain the acceleration in the appropriate direction. At each turn, the energy of the particle increases, and thus, the radius of the orbit gets larger and larger. After several turns, the particle reaches the rim of the dees and can hit a target where the collision occurs. Cyclotrons had limitations, for instance, the velocity of the accelerated particles could not exceed about $0.1c$ because the circular motion at high speed could not be maintained synchronous with the accelerating field.

Cyclotrons are now replaced by *synchrotrons*, where particles orbit in a fixed circular ring thanks to an adjustable magnetic field. The acceleration is realised thanks to a radio-frequency (RF) cavity located along the ring. RF cavities are metallic chambers that contain an electromagnetic field generated by an RF power generator. They are shaped so that electromagnetic waves become resonant inside the cavity. Their frequency is synchronised to accelerate the particles every time they pass the RF cavity. The Large Hadron Collider (LHC) at CERN (European Organisation for Nuclear Research) near Geneva, Switzerland, is nowadays the most emblematic and powerful synchrotron, colliding two beams of protons, each at an energy of 6.5 TeV, or for one month a year, beams of heavy ions with 2.56 TeV per nucleons. In the 2020s, it is expected that the proton beam energy would be upgraded to 13.6 or even 7 TeV (and the heavy ions beam to 2.76 TeV per nucleons). To accelerate particles to very high energy, synchrotrons require strong magnetic fields to maintain the particles on their circular path. At the LHC, magnetic fields up to 7.74 T are generated by 1 232 superconducting dipoles operating at 1.9 K, and eight superconducting RF cavities delivering 5 MV/m are used for the acceleration. Synchrotrons are limited by two main factors: the maximum magnetic field strength of the bending magnets and the *synchrotron radiation*. A charged particle in a circular orbit undergoes an acceleration and thus emits electromagnetic radiation, the synchrotron radiation, whose radiated power is proportional to $E^4/(m^4R^2)$, with E and m being the energy and mass of the particle, respectively, and R the radius of the orbit. To limit this radiation, it is thus better to use heavy particles and very large synchrotrons. That is why the LHC accelerates protons to 6.5 TeV in a ring about 27 km in circumference, whereas its predecessor, the Large Electron–Positron collider (LEP), accelerated electrons at most to 104.5 GeV in that same ring.

In order to circumvent synchrotron radiation, another strategy is to use *linear accelerators*. However, unlike circular accelerators, an RF cavity can accelerate particles only once on a linear trajectory. The acceleration voltage per unit length is thus the limiting factor in

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

the energy to which a linear accelerator can boost particles. In addition, beams can only collide once, limiting the probability of interaction. The Stanford Linear Collider (SLC), operating until 1998 at Stanford (CA, USA), was the longest linear accelerator (3.2 km), accelerating electrons and positrons to an energy of 50 GeV. It is, to date (and will stay even in the near future), the only example of a linear collider that has produced high energy particle collisions (i.e., tens of GeV) at a high rate.

In summary, particles can be produced from collisions between cosmic rays with the atmosphere or in a controlled environment using particle accelerators. Although accelerators are the primary tool of the high energy physicist, cosmic rays are also studied, particularly to understand the astrophysical phenomena that accelerate particles to such high energies, and for neutrino physics (see Chapter 9). In the history of particle physics, probing collisions of increasingly high energy has led to many discoveries. The higher the energy, the smaller the spatial extent of the object that can be probed. From this perspective, the key parameter of accelerators is the maximum energy reached by the accelerated particles. The other important parameter is the *luminosity* (defined in Section 3.5). The luminosity, $\mathcal{L}$, has the dimension of events per unit time per unit area and determines, for a given reaction, the number of potential collisions that can be produced per second, dN/dt. More specifically,

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \mathcal{L} \times \sigma, \tag{1.8}$$

where σ is the cross section of the reaction. The cross section is defined in Section 3.5, but at this stage, let us say that it is a quantity representative of the probability of the interaction. Therefore, the higher the luminosity, the smaller the cross section to be studied can be, revealing rare processes. With a luminosity of the order of 10^{34} cm^{-2} s^{-1}, the LHC yields about one billion proton–proton collisions per second at a centre-of-mass energy of 13 TeV. The integrated luminosity, i.e. the integral of the luminosity over the operating time of the collider, is then directly related to the total amount of collisions produced by the collider. It is, therefore, a metric of its performance.

1.4 Detection of Particles

The complexity of detectors has considerably increased over time, in particular, because of the ever higher particle energies involved and the very large number of particles produced in the collisions generated by modern colliders. Detectors now include devices able to identify particles and measure with a high precision their positions, momenta, energies and possibly their lifetimes. The precise description of the technology used in detectors goes beyond the scope of this book. There are dedicated books about this topic, such as those authored by Rossi (1952), Leo (1994) and Grupen and Shwartz (2011). This section is more focused on the interaction mechanisms exploited by particle physics detectors.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

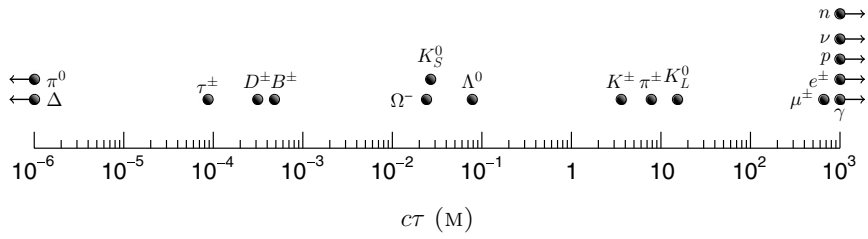


Fig. 1.2 Lifetime of particles multiplied by the speed of light. The arrows indicate that the actual value of $c\tau$ exceeds the scale in the figure.

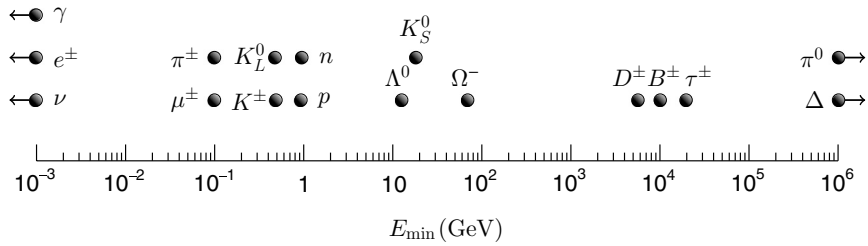


Fig. 1.3 Required minimum particle energy, $E_{\min}$ from Eq. (1.9) in GeV, to travel over a distance of 1 m. The arrows indicate that the actual value of $E_{\min}$ exceeds the scale in the figure.

1.4.1 General Considerations

Most of the particles are highly unstable with a lifetime so short that they almost immediately decay after their production. Hence, they cannot be observed directly in the detectors, and only their long-lived decay products can reach detectors or interact long enough with detectors. Figure 1.2 shows the average distance travelled in the vacuum by a representative sample of particles, assuming that their velocity is equivalent to the speed of light (ultra-relativistic approximation). Only a subset of unstable particles can travel over macroscopic distances, let us say at least a few millimetres, before decaying. All such particles decay by the weak interaction. If we require that a particle of mass m and lifetime τ travels on average over a distance l in the vacuum, the minimum energy is constrained to be (see Problem 1.4)

$$E_{\min} = mc^2 \sqrt{1 + \left(\frac{l}{c\tau}\right)^2}. \tag{1.9}$$

To give an order of magnitude, let us choose $l = 1$ m (the typical size of modern detectors is a few tens of metres). Figure 1.3 shows the corresponding values of $E_{\min}$ for various particles. Modern accelerators are powerful enough to produce collisions from which particles with energies of the order of a few GeV emerge. Therefore, we conclude from Fig. 1.3 that, statistically, the only particles that can possibly interact with detectors over a significant distance are:

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

$$\gamma, e^{\pm}, \nu\ (\bar{\nu}), \pi^{\pm}, \mu^{\pm}, K^{\pm}, K_S^0, K_L^0, p\ (\bar{p}), n\ (\bar{n}), \Lambda^0\ (\overline{\Lambda^0}). \tag{1.10}$$

Particles of this list can be classified into two categories: charged and uncharged particles. Since they experience different interactions with matter, different devices are needed for their detection.

1.4.2 Interaction of Charged Particles with Matter

Energy Loss by Ionisation

When a charged particle passes through a medium (called below *absorber*), it continuously interacts with the electrons of its atoms, or classically speaking, it undergoes inelastic collisions with the electrons. Thus, it transfers a fraction of its energy to the electrons of the absorber, ionising the atoms when the transfer is large enough or exciting them (raising the electrons to a higher lying shell) otherwise. The energy transfer causes an energy loss of the incident particle, thus reducing its velocity. The energy transfer to the electrons may be so large that these electrons may cause further ionisation of the atoms. Such electrons are called δ-rays. Scattering from nuclei also occurs. Since, in general, the mass of the incident particle is significantly smaller than that of the nucleus of the absorber (almost no recoil), very little energy is transferred by this mechanism. In this case, only elastic scattering occurs.[8]

The energy loss by ionisation or excitation is statistical in nature, governed by quantum mechanical probabilities. For a macroscopic path length, it occurs many times, and thus the fluctuations in the energy loss are small. As a first approximation, one can consider the average energy loss. For an incident particle with an electric charge ze, the mean energy loss by ionisation or excitation per unit length, $-\,\mathrm{d}E/\,\mathrm{d}x$, normalised to the absorber density ρ (g/cm^3) is given by the Bethe–Block formula (Bethe, 1930, 1932; Bloch, 1933)

$$\frac{1}{\rho}\left\langle -\frac{\mathrm{d}E}{\mathrm{d}x}\right\rangle_{\text{ion.}} \simeq Kz^2\frac{Z}{A}\frac{1}{\beta^2}\left(\ln\frac{2m_e c^2\gamma^2\beta^2}{I} - \beta^2 - \cdots\right). \tag{1.11}$$

This formula is an approximation valid for incident charged particles heavier than the electron, and for moderately relativistic velocities, $\beta\gamma \sim 0.1{-}1,000$, with $\beta = v/c$ and $\gamma = 1/\sqrt{1-\beta^2}$. In Eq. (1.11), $K = 0.3071$ MeV mol^{-1} cm^2 is a constant, Z is the atomic number of the absorber, A is its atomic mass in g mol^{-1} and m_e is the electron mass. The parameter I is called the mean excitation energy: it is an effective ionisation potential, averaged over all electrons of the absorber. It depends only on the absorber, and its value for various absorbers can be found in the literature (e.g., in Particle Data Group, 2022). To give an order of magnitude, a usual parametrisation for $z > 1$ is $I \sim 16\,Z^{0.9}$ eV (Grupen and Shwartz, 2011). The dots in Eq. (1.11) represent a set of small corrections. They concern incident particles moving at low velocities that are comparable to or smaller than those of atomic electrons but also high energy particles whose electric field is screened by the electric polarisation of the medium. They can be found in Particle Data Group (2022). The quantity $2m_e c^2\gamma^2\beta^2$ in the numerator of the logarithm in Eq. (1.11) is an

[8] By definition, in elastic scattering, the total kinetic energy is conserved.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

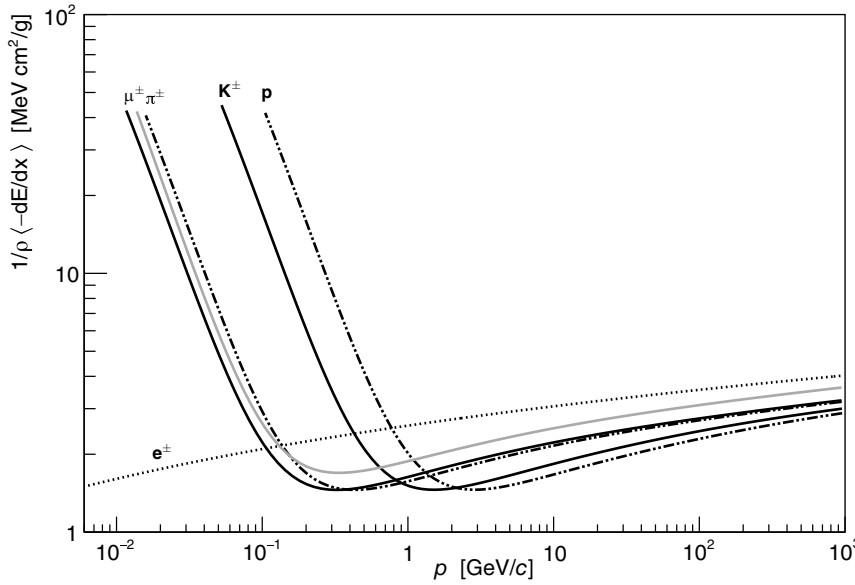


Fig. 1.4 The energy loss by ionisation or excitation (stopping power), $\langle \mathrm{d}E/\,\mathrm{d}x\rangle\,/\rho$, as a function of the momentum of different particles. The absorber here is copper ($Z = 29, A = 63.5\ \mathrm{g/mol}, I = 322\ \mathrm{eV}$), except for the grey curve, which represents the stopping power of muons in silicon ($Z = 14, A = 28.1\ \mathrm{g/mol}, I = 173\ \mathrm{eV}$).

approximation, valid at moderately relativistic velocities, of the maximum kinetic energy that can be transferred to an electron.[9]

For a given absorber, the energy loss in Eq. (1.11), also called the *stopping power*, depends only on the velocity (β) and the charge of the incident particle. It does not depend on its mass, to first order. It is shown in Fig. 1.4 as a function of the momentum $p = \gamma m\beta c$, for different particles passing through copper. The energy loss is maximal at low velocity since slower particles experience the electron field for a longer time. In this regime, $\mathrm{d}E/\mathrm{d}x$ is dominated by $1/\beta^2$ and decreases with increasing velocity. It reaches a minimum when $\gamma\beta = p/(mc) \sim 3{-}4$ or, equivalently, when $v \sim 0.95c{-}0.96c$. Particles at this point are called MIP or *Minimum Ionising Particles*. According to Fig. 1.4, an MIP loses about 1.4 MeV cm^2/g. This value varies little over a wide range of absorbers. At first order, the incidence of the absorber in Eq. (1.11) comes from the factor Z/A, which changes little for comparable Z. Hence, when $\mathrm{d}E/\,\mathrm{d}x$ is normalised to the density, as in Eq. (1.11), it is almost independent of the absorber. At higher energy than that of the MIP, the factor β becomes almost constant (equal to 1), and the factor in $1/\beta^2$ becomes irrelevant. The energy loss rises though, because of the term in $\ln(\gamma^2\beta^2)$, an effect called the relativistic rise. The previously mentioned corrections become significant, however, and tend to

[9] Without this simplification, the correct expression of the numerator in Eq. (1.11) is $\sqrt{2m_e c^2\gamma^2\beta^2 T_{\max}}$, with $T_{\max} = 2m_e c^2\gamma^2\beta^2/(1 + 2\gamma m_e/m + (m_e/m)^2)$, where m is the mass of the incident particle. Equation (1.11) is recovered when $2\gamma m_e/m \ll 1$ and $m \gg m_e$.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

moderate the rise. Hence, in practice, most relativistic particles have a mean energy loss by ionisation that is close to the minimum.

Formula (1.11) is valid for charged particles heavier than the electron. For electrons or positrons, one has to take into account that the mass of the incident particle and that of the target electron is the same [the derivation of Eq. (1.11) assumes that the incident particle remains undeflected during the collision]. In addition, incident electrons cannot be distinguished from those of the target, requiring an appropriate treatment in the calculation. The expression of the energy loss for electrons or positrons due to collisions on electrons can be found in Leo (1994) and Particle Data Group (2022). They are represented with the dashed line in Fig. 1.4, where they cannot be distinguished from one another in this momentum region. At relativistic energies, above tens of MeV, on the other hand, the energy loss of positrons and electrons is dominated by another source of energy loss: *bremsstrahlung*, i.e. the energy loss by radiation (see Section 'Energy Loss by Radiation: The Bremsstrahlung').

The $\mathrm{d}E/\,\mathrm{d}x$ curves allow the identification of charged particles. If both the momentum $p = \gamma\beta mc$ and the energy loss $\mathrm{d}E/\mathrm{d}x$ are measured with sufficient accuracy, it gives a point in the $(\mathrm{d}E/\,\mathrm{d}x, p)$ plane that can be used to determine the species of the particle.

Energy Loss by Radiation: The Bremsstrahlung

Charged particles interact not only with the electrons of atoms (contributing to the energy loss by ionisation and excitation of the atoms) but also with the Coulomb field of the nuclei. The incident particle is then deviated from its straight-line course by the electrical interaction with the nuclei. It contributes to *multiple scattering* but also to the emission of electromagnetic radiation due to the acceleration of the charged particle. As synchrotron radiation is emitted because of the acceleration in a magnetic field, *bremsstrahlung* is radiation emitted because of the acceleration (or rather a deceleration) in the electric Coulomb field (the German word bremsstrahlung means 'deceleration radiation').

For high energies, the energy loss by bremsstrahlung can be described by Rossi (1952)

$$\left\langle -\frac{\mathrm{d}E}{\mathrm{d}x}\right\rangle_{\text{brem.}} \simeq 4\alpha N_A \frac{Z^2}{A}\left(\frac{1}{4\pi\epsilon_0}\frac{(ze)^2}{mc^2}\right) E \ln\left(183\, Z^{-1/3}\right), \tag{1.12}$$

where ze, m and E are the charge, mass and energy of the incident charged particle, respectively, and Z and A are, respectively, the atomic number and atomic mass of the absorber. Whereas the energy loss by ionisation varies logarithmically with energy (Fig. 1.4) and linearly with Z [Eq. (1.11)], the energy loss by bremsstrahlung increases linearly with E and quadratically with Z. In addition, it is proportional to the inverse of the particle mass squared. Hence, at high energy, only low mass particles, namely positrons and electrons, lose significant energy by bremsstrahlung. The *critical energy*, E_c, is defined as

$$\left\langle -\frac{\mathrm{d}E}{\mathrm{d}x}\right\rangle_{\text{brem.}} = \left\langle -\frac{\mathrm{d}E}{\mathrm{d}x}\right\rangle_{\text{ion.}} \quad \text{for } E = E_c. \tag{1.13}$$

For electrons and positrons, it is about a few tens MeV in most solid absorbers, with an approximate formula being $E_c \simeq 1600\, m_e c^2/Z$ (Leo, 1994). For the next lightest charged particle, the muon, the energy loss by bremsstrahlung dominates over that by ionisation for

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

muon energies of a few hundred GeV (Lohmann et al., 1985).[10] With such a high value, we will restrict ourselves to the bremsstrahlung of electrons and positrons.

Since the energy loss by bremsstrahlung is proportional to energy, Eq. (1.12) can be rewritten (forgetting the mean value) as

$$-\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{E}{X_0},$$

whose solution is

$$E(x) = E_0 \exp(-x/X_0). \tag{1.14}$$

Here, the quantity X_0 is called the *radiation length*. It corresponds to the average length over which the energy is reduced by a factor $1/e$, due to bremsstrahlung radiation. According to Eq. (1.12), for electrons and positrons $z^2 = 1$ and $m = m_e$, such that introducing the classical radius of the electron $r_e = e^2/(4\pi\epsilon_0 m_e c^2) = 2.82 \times 10^{-13}$ cm, the radiation length in g/cm^2 reads

$$X_0 = \frac{A}{4\alpha N_A Z^2 r_e^2 \ln\left(183\, Z^{-1/3}\right)}. \tag{1.15}$$

Table 1.6 gives the values of the critical energy and the radiation length for a few common absorbers. Formula (1.15) is approximate: it neglects the bremsstrahlung due to the electron cloud and possible screening of the Coulomb field of nuclei by electrons. Therefore, the values in Table 1.6 differ from the results of Eq. (1.15) by about 10−20%. A better approximate formula is (Grupen and Shwartz, 2011)

$$X_0 = \frac{716.4\, A\, [\mathrm{g/mol}]}{Z(Z+1)\ln\left(287\sqrt{Z}\right)}\ \mathrm{g/cm^2}. \tag{1.16}$$

Since the energy loss by bremsstrahlung is proportional to Z^2 and thus X_0 in $1/Z^2$ at first order, if one wants to stop a high energy electron, it is better to use absorbers with a high atomic number. This determines the choice of the materials used in electromagnetic calorimeters, as we will see in Section 1.4.4.

Emission of Cherenkov Radiation

Another kind of radiation occurs (but not in the same wavelength spectrum) when a charged particle traverses a medium at a velocity larger than the phase velocity of light in that medium, i.e. when the velocity of the particle satisfies

$$v > c/n,$$

where n is the index of refraction. The passage of the charged particle transiently polarises the atoms or the molecules of the medium, generating electric dipoles. Since the polarisation varies with time, it is well known from classical electrodynamics that the dipoles emit

[10] Actually, for high energy muons, another radiation process called *pair production* dominates, where the interaction of a muon with a nucleus produces $\mu^{\pm} + e^+ + e^- + X$, with X being any other decay product. It is about 1.5 times larger than bremsstrahlung at energies above a few hundred GeV (Lohmann et al., 1985). Hence, the energy loss by ionisation becomes negligible with respect to all other contributions for $E \gtrsim 350$ GeV in copper (Particle Data Group, 2022).

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

Table 1.6 Critical energies (electron) and radiation lengths for various absorbers.

Absorber	Z	A	E_c	X_0	
		(g/mol)	(MeV)	(g/cm^2)	(cm)
Air (dry, 1 atm)	–	–	87.92	36.62	30390
H_2O	–	–	78.33	36.08	36.08
Pb	82	207.2	7.43	6.37	0.56
Cu	29	63.55	19.42	12.86	1.44
Fe	26	55.85	21.68	13.84	1.76

Source: From Particle Data Group (2022, Section: 'Atomic and Nuclear Properties of Materials').

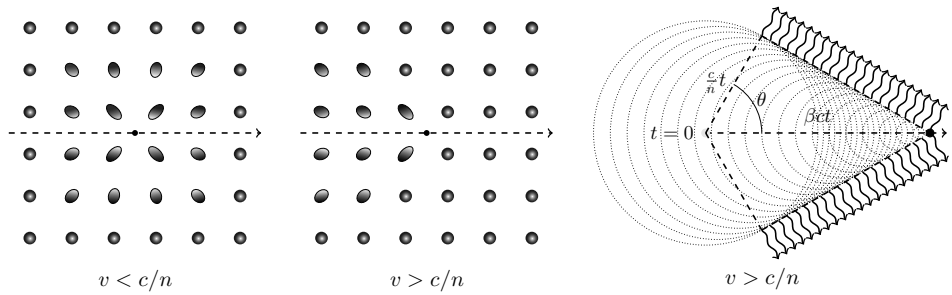


Fig. 1.5 Illustration of the Cherenkov effect when an electron (black circle) passes through a medium. Left-hand and middle diagrams: polarisation of the medium when $v < c/n$ and $v > c/n$, respectively. Atoms are represented by shaded circles or ellipses. The shading is white for positive charges and black for negative charges. Right-hand side: emission of Cherenkov light (wavy lines) with its characteristic angle θ. The dotted circles represent the wavefront of the light at different times.

electromagnetic radiation (see, e.g., Jackson, 1998). The emission stops when the atoms or the molecules return to their initial state without polarisation. However, when dipoles are symmetrically arranged with respect to the track of the charged particle, the resulting electric field of all dipoles vanishes, preventing the emission of the radiation. This is only possible when the particle travels slower than light in the medium, $v < c/n$, as illustrated in the left-hand diagram of Fig. 1.5. In contrast, for velocities greater than c/n (middle diagram), there is a resultant dipole field along the axis of the particle track (and not elsewhere), varying with time. Hence, radiation is emitted at each point along the axis. The wavelets from all portions of the axis are then in phase with one another (right-hand diagram of Fig. 1.5), such that a distant observer can see the Cherenkov light only at a single angle, θ, with respect to the track of the particle. By symmetry about the axis of the particle, the light emitted on each point of the axis propagates along the surface of a cone. Elementary geometry allows us to determine the opening angle of the cone. Between $t = 0$ and t, the charged particle travels over a distance $vt = \beta ct$, while the light emitted at $t = 0$ defines a wavefront of radius $c/n \times t$. Thus, the cosine of the angle, the ratio of these two quantities, satisfies

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

$$\cos\theta = \frac{1}{\beta n}. \tag{1.17}$$

For example, in water where $n = 1.33$, the maximum Cherenkov angle is reached when $\beta = 1$, giving $\theta = 41.4^\circ$, while in air ($n = 1.000293$), it is only $\theta = 1.39^\circ$.

One can show that the number of photons radiated per unit path per unit wavelength is [see Jelley (1958), an excellent book on the Cherenkov effect and its applications]

$$\frac{\mathrm{d}^2N}{\mathrm{d}x\,\mathrm{d}\lambda} = \frac{2\pi\alpha z^2}{\lambda^2}\sin^2\theta = \frac{2\pi\alpha z^2}{\lambda^2}\left(1 - \frac{1}{\beta^2 n^2}\right), \tag{1.18}$$

where the charge of the particle is ze. A naive interpretation of Eq. (1.18) would lead us to conclude that very short wavelengths should be favoured. However, all media are more or less dispersive (except the vacuum), meaning that the refractive index actually depends on the wavelength of the photon (or its energy), $n = n(\lambda)$. In the X-ray region and at higher energy, $n(\lambda)$ is always less than 1,[11] forbidding the emission of radiation since $\beta = v/c > n > 1$ is impossible. Hence, ultraviolet and visible photons are the most numerous. This imposes the use of a transparent medium for the construction of detectors using the Cherenkov effect.

The energy loss due to the emission of Cherenkov photons is negligible compared with the ionisation or bremsstrahlung loss. Since $E_\gamma = hc/\lambda$, it follows from Eq. (1.18),

$$\frac{\mathrm{d}^2N}{\mathrm{d}x\,\mathrm{d}E_\gamma} = \frac{\alpha z^2}{\hbar c}\sin^2\theta \simeq 370\, z^2 \sin^2\theta\ \mathrm{eV}^{-1}\mathrm{cm}^{-1}. \tag{1.19}$$

Thus, the energy lost by the charged particle is $-d^2E = E_\gamma d^2N$, yielding

$$-\left.\frac{\mathrm{d}E}{\mathrm{d}x}\right|_{\mathrm{Cher.}} = \frac{\alpha z^2}{\hbar c}\int_{\beta n>1}\sin^2\theta\, E_\gamma\,\mathrm{d}E_\gamma \simeq 370\, z^2 \int_{\beta n>1}\sin^2\theta\, E_\gamma\,\mathrm{d}E_\gamma. \tag{1.20}$$

Assuming that the medium is reasonably not dispersive between 0 and $E_\gamma^{\max}$, the integration of Eq. (1.20) gives $370\, z^2 \sin^2\theta\ (E_\gamma^{\max})^2/2$. For instance, with $E_\gamma^{\max} = 6.6$ eV (ultraviolet photon, $\omega \sim 10^{16}$ Hz), a 100 MeV electron ($z^2 = 1, \beta \sim 1$) passing through 1 cm of water ($\sin^2\theta = 0.25$) loses about 2×10^{-3} MeV energy. That same electron, according to Eq. (1.14) and Table 1.6, would lose about 2.7 MeV energy by bremsstrahlung– three orders of magnitude larger. Note that the energy loss due to Cherenkov radiation is already included in Eq. (1.11) via the correction factor at high energy that takes into account the polarisation of the medium. This is the part preventing $-\,\mathrm{d}E/\mathrm{d}x$ to rise as $\log(\gamma\beta)$ at high energy (Fermi plateau). It is important to realise that both the Cherenkov radiation and bremsstrahlung radiation are of different nature: the Cherenkov radiation arises only from the macroscopic properties of the medium, whereas in the case of bremsstrahlung radiation, it is the individual interaction of the charged particle with the Coulomb field of atoms that matters.

The measurement of the Cherenkov angle provides another way of identifying particles, as soon as the momentum p is known, since from Eq. (1.17),

$$\theta = \arccos\left(\frac{E}{npc}\right) = \arccos\left(\frac{\sqrt{p^2c^2 + m^2c^4}}{npc}\right). \tag{1.21}$$

[11] The phase velocity of light can be greater than c.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

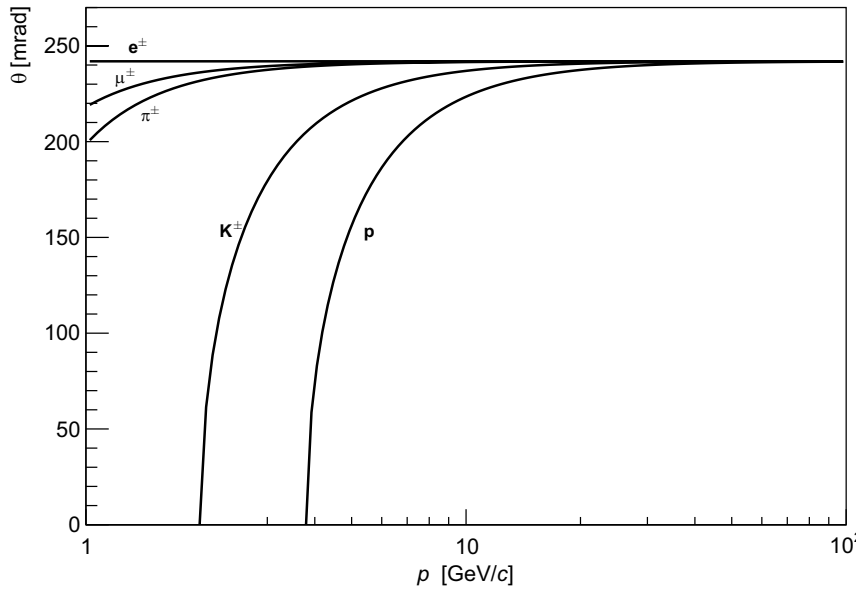


Fig. 1.6 Cherenkov angle as a function of the momentum of various particles. The medium here is the aerogel, with a reflective index $n = 1.03$.

Identification of particles is only possible when the Cherenkov effect is present, i.e. when $\beta > 1/n$. It imposes the following constraints on the energy and momentum of the particle:

$$E > \frac{n}{\sqrt{n^2-1}}mc^2, \quad p > \frac{1}{\sqrt{n^2-1}}mc. \tag{1.22}$$

Figure 1.6 shows the Cherenkov angles of various particles as a function of their momentum in the aerogel, a mixture of SiO_2 and H_2O. The region of momenta where identification is possible is narrow; thus, detectors frequently use several media to extend it.

For completeness, radiation similar to the Cherenkov effect exists when a charged particle crosses from one medium to another with different dielectric properties, for example, from the vacuum to a pure dielectric or a perfect conductor. This *transition radiation* is emitted at the interface between the two media. Unlike Cherenkov radiation, it occurs at any velocity of the particle. However, the probability increases with γ, so in practice, only highly relativistic particles ($\gamma \gtrsim 10^3$) emit substantial radiation. This characteristic is sometimes used in detectors to discriminate heavy particles, such as hadrons from lighter ones, typically electrons, since for a given energy $E = \gamma mc^2$, they have very different γ factors. The interested reader can consult Grupen and Shwartz (2011) where many examples of transition radiation detectors are given.

1.4.3 Interaction of Neutral Particles with Matter

Uncharged particles do not interact through the Coulomb force, and therefore, while passing through a medium, they must first undergo a strong, weak or, in the case of photons,

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

electromagnetic interaction. The first interaction of the particle with the medium often involves the nucleus of constituent atoms of the medium and produces charged particles that are then detected.

Specific Interactions of High-Energy Photons

Interactions of high energy photon (X-ray or γ-ray) produce electrons through three main modalities: the *photoelectric effect*, *Compton scattering* and *pair production* of electrons.

In the photoelectric effect, a photon is absorbed by an atomic electron, leading to the ejection of the electron from the atom: $\gamma + \text{atom} \rightarrow \text{atom}^+ + e^-$. This process is the predominant mode of interaction for γ-rays or X-rays of relatively low energy, typically below 0.5 MeV. The ejected electron, called *photoelectron*, has energy given by $E_{\text{p.e}} = h\nu - E_b$, where $h\nu$ is the energy of the incident photon and E_b is the binding energy of the photoelectron in its original atomic shell (the recoil of the atoms can be safely ignored). Since E_b is small compared with γ-ray energy, the photoelectron carries most of the photon energy. Hence, for an evaluation of the photon energy, this process is ideal.

In Compton scattering, the incoming photon is not absorbed but deflected by an electron, transferring a portion of its energy to the electron that recoils (to conserve the energy and momentum): $\gamma + e^- \rightarrow \gamma + e^-$. The energy spectrum of the recoil electron is then a continuum, whose maximum energy is $m_e c^2 + E_\gamma \times \epsilon/(1+\epsilon)$, with $\epsilon = 2E_\gamma/(m_e c^2)$. In Chapter 6, we will learn how to calculate the cross section of this process, assuming that the electron is free. However, as soon as the photon energy is much larger than the binding energy of the electron, which is the case for γ-ray photons, the results obtained in Chapter 6 remain valid (including the maximum value given above). Compton scattering is the dominant process for photon energy between 0.5 and 10 MeV.

Finally, at energies higher than 10 MeV, pair production dominates. It mostly results from the interaction of a photon with the Coulomb field of the nucleus: $\gamma + \text{nucleus} \rightarrow \text{nucleus} + e^+ + e^-$. In this process, the photon is thus absorbed as in the photoelectric effect. Pair production can be viewed as a process similar to bremsstrahlung since for the latter, $e^- + \text{nucleus} \rightarrow \text{nucleus} + e^- + \gamma$. Both bremsstrahlung and pair production involve the same set of particles, provided an electron is exchanged for a positron. We will learn in Chapter 6 that, because of quantum field theory properties, the two processes can then be calculated from one another. Consequently, just as the energy of electrons decreases exponentially in matter by bremsstrahlung emission as a function of the radiation length, X_0 [cf. Eq. (1.14)], for pair production, the intensity of high energy photons behaves the same way, i.e.

$$I(x) = I_0\, e^{-x/\lambda}, \tag{1.23}$$

where, for photon energy greater than 1 GeV, the *attenuation length* or *mean free path* λ is given by Particle Data Group (2022)

$$\lambda \simeq \frac{9}{7} X_0. \tag{1.24}$$

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

The parameter $\mu = 1/\lambda$ is thus an absorption coefficient. When it is normalised to the density of the medium $\mu' = \mu/\rho$ (like the radiation length), it is called the *mass absorption coefficient*.

Interaction of Hadrons

Hadrons experience the strong interaction and can interact with the nuclei of the medium. For neutral hadrons, such as neutrons, this is the main interaction in matter. Since the strong force is a short-range interaction, the probability that the hadron is close enough to a given nucleus (about 1 fm) is low, and hence neutral hadrons are penetrating particles. Charged hadrons can also have electromagnetic interactions, leading to the mechanisms of energy loss already presented.

The interactions of neutrons depend strongly on their energy. At low kinetic energy, in the MeV region and below, elastic scattering of neutrons with nuclei contributes to their energy loss. Inelastic scattering is also possible, leaving the nucleus in an excited state that will further de-excite emitting photons or charged particles. When neutrons are sufficiently slowed down, they may be captured by a nucleus that emits photons, or they may induce a nuclear reaction (fission etc.). In particle physics, those low-energy neutrons are generally identified via those secondary particles (photons or charged particles). On the other hand, neutrons or any hadrons with high energy (kinetic energy above 100 MeV, so most of the hadrons produced with accelerators) undergo nuclear reactions that produce secondary hadrons, that themselves generate nuclear reactions and so on. This process produces a *hadronic shower* that can be captured by a hadronic calorimeter (see Section 'Detection of Hadrons: Hadronic Calorimeters').

1.4.4 Detectors

Measurement of the Momentum and dE/dx: Tracker Detectors

Track detectors, or *trackers* for short, are detectors that measure the particle tracks (i.e., the trajectory of the particle and the energy deposited along the track). Because of the large number of particles produced by collisions at modern colliders, most detectors measuring p and dE/dx are nowadays based on semiconductors devices (mainly silicon). The previous generation of detectors was often gaseous detectors like *time projection chambers*, where the passage of the ionising particle created electron–ion pairs collected by an electric field. In the case of semiconductors, the electric charge is carried by electrons that have been excited to the conduction band from the valence band and by the corresponding holes created in the valence band.[12] The crystal lattice of semiconductors is generally doped with elements that can easily provide an electron (donor) or capture it (acceptor). A few impurity atoms per billion semiconductor atoms are sufficient. Silicon has four valence electrons. Donors then have five valence electrons, with the excess electron becoming easily a free electron (its energy level is very close to the lower bound of the conduction band, about

[12] The reader not familiar with solid state physics can consult, for example, Kittel (2005).

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

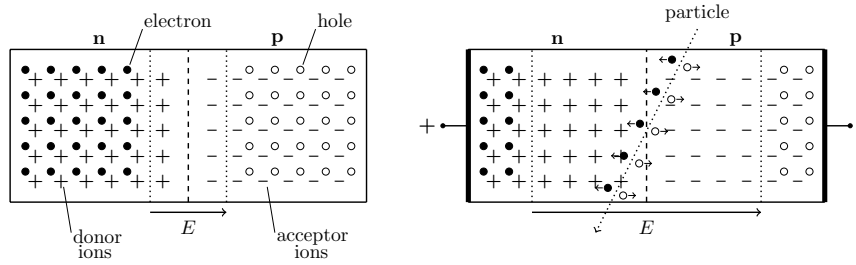


Fig. 1.7 Working principle of a p–n junction. On the left-hand side, no voltage is applied. On the right-hand side, a reverse bias is applied to the junction i.e. a positive voltage on the n-side and a negative voltage on the p-side. The effect of the passage of a particle is illustrated with the drift of the charge carriers.

0.05 eV in silicon). Silicon doped in this way is called n-doped silicon (n for negative). If acceptors with three valence electrons are added to the crystal lattice, one electron is missing for the covalent bond, creating a hole (and a new energy level in the energy gap, very close to the upper bound of the valence band). It forms p-doped silicon. Silicon detectors are based on p–n junctions, where p-doped silicon and n-doped silicon are in contact, as illustrated in the left-hand sketch in Fig. 1.7. In the contact area, there is a natural migration of electrons from the n-region to the p-region to annihilate the holes. Similarly, holes from the p-region diffuse towards the n-region. Hence, both sides of the junction are filled with ionised donors and acceptors, which are immobile, as they are bound in the crystal lattice. This creates an electric field, which balances the diffusion force of electrons and holes, generating a depleted region of free charges (electrons or holes). When an ionising particle enters the depleted region, electron–hole pairs are created. Both charge carriers then feel electrostatic forces and drift in opposite directions. These charges can then be collected by electrodes, inducing a current pulse. However, the intrinsic electric field is not intense enough to collect enough charges. Moreover, the depletion region is too thin to allow high energy particles to lose enough energy. Therefore, a voltage is applied to the junction (right-hand sketch in Fig. 1.7) such that it increases the electric field and the thickness of the depleted region. The total charge collected is proportional to the energy deposited in the depletion layer. The average energy required for the creation of an electron–hole pair is typically about 3 eV, which is an order of magnitude less than the energy required for the creation of electron–ion pairs in a gaseous detector.[13] Hence, for a given energy, semiconductors produce about 10 times more electron–hole pairs than electron–ion pairs. Moreover, they have a greater density and thus a larger stopping power. This explains their popularity despite their relatively high cost.

Trackers based on silicon use a very large number of silicon sensors with p–n junctions (some have nearly 100 million) that can be shaped into thin strips or pixels. In its simplest form, a silicon strip detector is constructed as an array of p-type junctions on a single n-type

[13] More specifically, for silicon, it is 3.62 eV at room temperature. Given that the energy gap is 1.14 eV, only one-third of the energy is spent on the production of the electron–hole pair. The other two-thirds excite the vibration states of the lattice.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

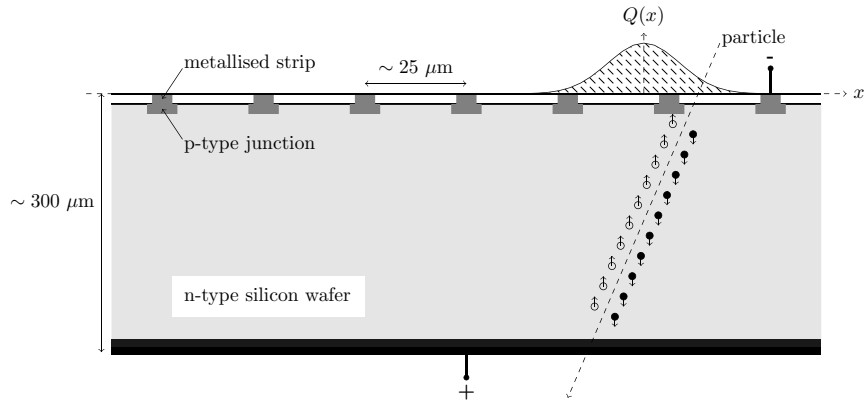


Fig. 1.8 Simplified schematic layout of a silicon strip detector. Electrons are represented by the black circles and holes by the open circles. The quantity $Q(x)$ represents the charge collected over several strips. Weighting the position of the strips by their charge gives interpolation of the particle position x.

silicon wafer (Fig. 1.8). The typical strip width is 10 μm on an about 25 μm pitch. When a particle passes through the depleted region (most of the n-type region is usually depleted), electron–hole pairs are produced, inducing a current in the strips, collected by readout amplifiers (one per strip). Given that an MIP leaves about 1.66 MeV/cm^2 (see the grey curve in Fig. 1.4), when a particle crosses 300 μm of silicon (density $\rho = 2.33$ g/cm^3), it loses about 116 keV. With 3.62 eV required to create a pair, we conclude that about 30 000 pairs are produced, giving a clear signal. The distribution of the collected charges over multiple strips allows the position interpolation of the particle in one dimension. To give an order of magnitude, the position resolution is about a few micrometres.[14] Double-sided strips with two orthogonal planes of silicon strips as well as an array of silicon pixels give position information in two dimensions. Detectors installed at accelerators have several concentric layers of silicon strips or pixels surrounding the collision region. They provide the third dimension of position reconstruction.

Track detectors frequently operate in a strong magnetic field $\boldsymbol{B}$ that deflects the charged particle trajectory by virtue of the Lorentz force,

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = q\boldsymbol{v} \times \boldsymbol{B}, \tag{1.25}$$

where $\boldsymbol{p} = \gamma m\boldsymbol{v}$ and q is the charge of the particle. Solving this equation (see Problem 1.5) shows that the trajectory is that of non-relativistic mechanics, i.e. a helical trajectory, except that the mass of the particle, m, is substituted with γm,

$$\begin{cases} x(t) = & x(0) + [p_y(0) + p_x(0)\sin(\omega t) - p_y(0)\cos(\omega t)]/(q|\boldsymbol{B}|) \\ y(t) = & y(0) + [-p_x(0) + p_y(0)\sin(\omega t) + p_x(0)\cos(\omega t)]/(q|\boldsymbol{B}|) \\ z(t) = & z(0) + p_z(0)\,\omega t/(q|\boldsymbol{B}|), \end{cases} \tag{1.26}$$

[14] If we assume that a single strip is read out with a pitch p between strips, and that the track could be anywhere between $-p/2$ and $p/2$ with a uniform distribution, simple statistics show that the position resolution is $p/\sqrt{12}$. With $p = 25$ μm, we obtain 7 μm.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

where $\omega = q|\boldsymbol{B}|/(\gamma m)$ and $\boldsymbol{B}$ is chosen along z. Projected on the plane orthogonal to the magnetic field, (x, y), the trajectory is a circle with a bending radius R satisfying

$$R = \left| \frac{\boldsymbol{p}_\perp}{q\boldsymbol{B}} \right|, \tag{1.27}$$

where $|\boldsymbol{p}_\perp| = \sqrt{p_x^2 + p_y^2}$ is the transversal component of the momentum. Notice that $|\boldsymbol{p}|$, $|\boldsymbol{p}_\perp|$ and p_z are constants of the motion [derive Eq. (1.26) to check it]. Therefore, we can introduce λ, often called the *pitch angle* of the helix, the angle between the projected track tangent in the (x, y) plane and the track tangent,

$$\cos\lambda = \left| \frac{\boldsymbol{p}_\perp}{\boldsymbol{p}} \right|, \quad \tan\lambda = \frac{p_z}{|\boldsymbol{p}_\perp|}. \tag{1.28}$$

Similarly, let us introduce ϕ, the angle between the x-axis and the projected track tangent at $t = 0$, i.e.

$$\cos\phi = \frac{p_x(0)}{|\boldsymbol{p}_\perp|}, \ \sin\phi = \frac{p_y(0)}{|\boldsymbol{p}_\perp|}.$$

Using these two angles and Eq. (1.27) yields the simple trajectory equations

$$\begin{cases} x(t) = & x(0) + R\sin\phi + R\sin(\omega t - \phi) \\ y(t) = & y(0) - R\cos\phi + R\cos(\omega t - \phi) \\ z(t) = & z(0) + R\,\omega t \tan\lambda. \end{cases} \tag{1.29}$$

The quantity $R\omega t$ represents the arc length of the track circle in the transverse plane. Figure 1.9 illustrates how the measurement of the signals in several layers of the tracker gives access to the bending radius in the transverse plane and the arc length (left-hand schema).

The angle λ can then be evaluated by fitting the slope in the plane (arc length, z) as shown in the right-hand schema. Having R and λ, we can then access the total momentum thanks to Eqs. (1.27) and (1.28). Using units where the momentum is in GeV/c and the charge is in units of the elementary charge e yields

$$|\boldsymbol{p}|\cos\lambda = \delta\,|q\boldsymbol{B}|\,R, \tag{1.30}$$

where $\delta = 10^{-9}c \simeq 0.3$. One can show [see Grupen and Shwartz (2011), for instance] that the relative resolution of the momentum measurement can be parametrised as

$$\frac{\sigma(|\boldsymbol{p}|)}{|\boldsymbol{p}|} = \sqrt{a^2|\boldsymbol{p}|^2 + b^2}, \tag{1.31}$$

where a and b are two constants encoding the intrinsic resolution coming from the measurement points along the track and the effect of multiple scattering, respectively. Qualitatively, the linear dependence of the first term with the momentum can be understood as follows: when the momentum is very high, the bending radius becomes almost a straight line, and thus, the uncertainty on the momentum gets very large. This follows from the fact that what is measured is not directly the bending radius but rather its inverse, the curvature $\kappa = 1/R$ (see Problem 1.6). Since according to Eq. (1.30) the uncertainty of κ

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

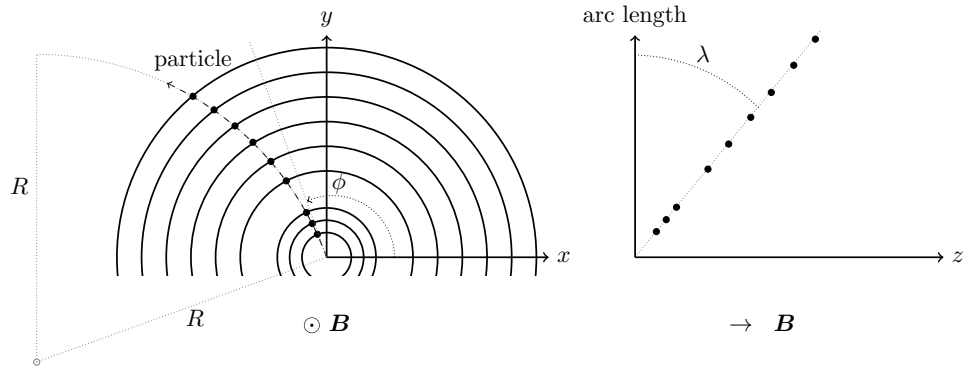


Fig. 1.9 A tracker with nine detecting layers. The trajectory, given by Eq. (1.29), is depicted with the dashed line. The position measurements on layers are represented by black circles. They enable the measurement of R, ϕ and λ and thus the momentum. In this example, the particle is produced in the centre of the detector. Otherwise, the distance of the closest approach to the origin in the (x, y) plane would have to be determined.

is $\sigma(\kappa) \propto \sigma(1/|\boldsymbol{p}|) = \sigma(|\boldsymbol{p}|)/|\boldsymbol{p}|^2$, it follows that $\sigma(|\boldsymbol{p}|)/|\boldsymbol{p}| \propto |\boldsymbol{p}|$. A typical order of magnitude of the constant a in Eq. (1.31) is $10^{-3} - 10^{-4}$, when $|\boldsymbol{p}|$ is expressed in GeV/c.

Detection of High-Energy Electrons and Photons: Electromagnetic Calorimeters

Calorimeters are detectors that aim to measure the energy of a particle, neutral or charged, by converting its energy into a measurable signal. Ideally, a calorimeter absorbs the whole energy of a particle. Their role is complementary to that of trackers that measure the momentum of a particle, ideally without perturbing the particle track much.

Electromagnetic calorimeters are dedicated to the energy measurement of photons and electrons. They are optimised to exploit the energy-loss mechanisms of these particles at high energies, in particular, bremsstrahlung emission for electrons and pair production for photons (see Sections 1.4.2 and 1.4.3). Via the interaction with the absorber of the calorimeter, a high energy electron emits a bremsstrahlung photon, which creates an electron–positron pair. The members of the pair can in turn emit photons, and the process continues until the produced particles have energies below the critical energy, at which point they lose their energy by ionisation or Compton scattering, for electrons and photons, respectively. This process creates a cascade of electrons, positrons and photons, called an *electromagnetic shower*. Similarly, when a high energy photon enters the calorimeter, it converts first into an electron–positron pair that generates the electromagnetic shower. An illustration is given in Fig. 1.1, p. 13, where the two photons that are the decay products of the π^0 on the left-hand side of the sketch are the source of the electromagnetic shower in the air. Air is not a good calorimeter, with its radiation length, X_0, being too large (see Table 1.6). It is much better to use dense materials with high atomic numbers, favouring large energy loss, such that the value of X_0 is small [cf. Eq. (1.15)].

Although the development of the electromagnetic shower is a statistical process, its main characteristics can be inferred from a simple model (proposed by Heitler), where after each

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

radiation length, X_0, the number of particles is doubled and their energy is divided by 2. In other words, in the processes $\gamma \to e^+ + e^-$ or $e^\pm \to \gamma + e^\pm$, the energy of the initial particle is assumed to be symmetrically shared between the particles of the final state. At a depth of t radiation lengths, the number of particles is then $N(t) = 2^t$, and their mean energy $E(t) = E_0/2^t$, where E_0 is the energy of the incident particle in the calorimeter. The shower should stop its development when the energy of the particles is below the critical energy, $E(t) < E_c$. The maximum number of particles is then reached after $t_{\max}$ radiation lengths, with $E(t_{\max}) = E_c$, i.e.

$$t_{\max} = \frac{\ln(E_0/E_c)}{\ln(2)}, \quad N(t_{\max}) = \frac{E_0}{E_c}. \tag{1.32}$$

Hence, after an exponential rise of the number of particles, the development of an electromagnetic shower reaches a maximum at a depth (in units of the radiation length) that scales logarithmically with the initial energy E_0, while the number of particles is proportional to E_0. For instance, according to Table 1.6, a 100 GeV electromagnetic shower reaches its maximum at $t_{\max}X_0 \simeq 12 \times 1.76 = 21$ cm in iron and about $14 \times 0.56 = 8$ cm in lead. A calorimeter built with a lead absorber is thus much more compact.

Equation (1.32) gives information only about the longitudinal profile (i.e., along the direction of the incident particle) of the shower. However, the shower also expands transversally, in particular, because multiple scattering of electrons becomes significant below the critical energy. The transversal profile is characterised by a lateral width, R_M, called the Molière *radius*, which is given by Particle Data Group (2022)

$$R_M = \frac{\sqrt{4\pi\alpha} m_e c^2}{E_c} X_0 \simeq \frac{21.2\ [\text{MeV}]}{E_c} X_0. \tag{1.33}$$

The Molière radius is such that, on average, 90% of the total energy of the shower (and thus, of the incident particle, approximately) is contained in a cylinder around the shower axis whose radius is $r = R_M$, and 95% for $r = 2R_M$.

An accurate description of showers based on analytical formulas is complicated. They are, however, rather well described by Monte-Carlo simulation.[15] A simulation of an electromagnetic shower created by a 100 GeV photon in the air is shown in Fig. 1.10a. The shape in denser materials is similar, with a much smaller longitudinal and lateral extension.

Calorimeters can be classified into two categories: homogeneous calorimeters and sampling calorimeters. In the first one, the same material combines the properties of an absorber and the ability to produce a measurable signal. An example is the Compact Muon Solenoid (CMS) electromagnetic calorimeter, made of 61 200 lead tungstate ($PbWO_4$) crystals in the central region of the detector (the so-called barrel). These crystals are inorganic scintillators,[16] with a length of 25 radiation lengths and a transverse size at the front face of about $R_M \times R_M$, ($X_0 = 0.85$ cm and $R_M = 2.19$ cm in $PbWO_4$ crystals). An electromagnetic shower is then fully contained in length and spreads over a few crystals

[15] Monte carlo simulations use random numbers to solve problems with many degrees of freedom.

[16] A scintillator is a medium that converts the excitation of its constituents (for example, the crystal lattice), due to the energy loss of charged particles, into visible (or often ultraviolet) light. There are many types of scintillators: organic crystals, glass, or various other liquid, plastic or gaseous materials etc. Plastic scintillators are often used because they can be easily shaped. See Leo (1994) for more details.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

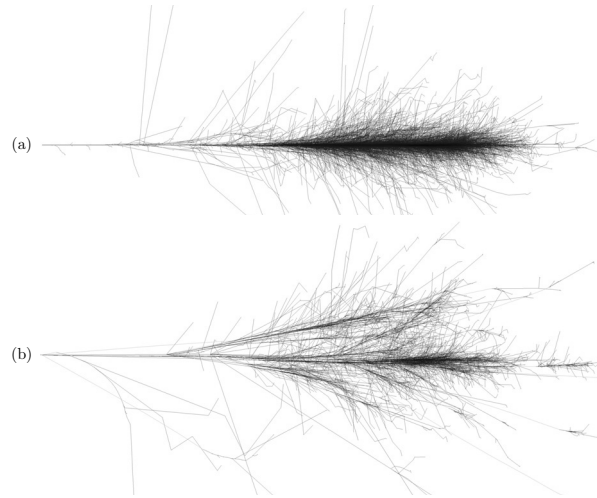


Fig. 1.10 Simulation of showers produced by a 100 GeV photon (a) and a 100 GeV proton (b) in the air. The longitudinal size is 30.1 km and the transversal size, 10 km. Images from F. Schmidt, J. Knapp, CORSIKA Shower Images, 2005, www-zeuthen.desy.de/jknapp/fs/showerimages.html.

transversally. Due to the large magnetic field (4 T), conventional photomultipliers (presented in the next section) cannot be used to detect the scintillation photons (their gain and linearity are affected by the field). Instead, avalanche photodiodes (denoted APD) are used to read the scintillation light. They are semiconductor devices (cf. Fig. 1.7) configured with a high reverse bias voltage and optimised in terms of doping layers to reproduce the avalanche of conventional photomultipliers when a photon enters the device (see the next section).

The second category of calorimeters are sampling calorimeters that alternate the layers of the absorber material with the active media that provide the measurable signal. Hence, only a sample of the total energy deposition is measured, limiting their energy resolution. However, they generally have the advantage of being more economical.

The relative energy resolution $\sigma(E)/E$ of calorimeters improves with the energy of the incident particle because the higher the energy, the larger the number of particles that are statistically produced in the shower. A standard parametrisation is

$$\frac{\sigma(E)}{E} = \sqrt{\left(\frac{a}{\sqrt{E}}\right)^2 + \left(\frac{b}{E}\right)^2 + c^2}, \tag{1.34}$$

where a, b and c are constants reflecting the different sources contributing to the resolution. For showers with energies a few tens of GeV, the dominating term is the stochastic term, $a/\sqrt{E}$, due to the fluctuations related to the physical development of the shower. At very high energy (hundreds of GeV and beyond), the limiting factor is the constant term, c (generally a fraction of percent), which includes instrumental contributions that do not

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

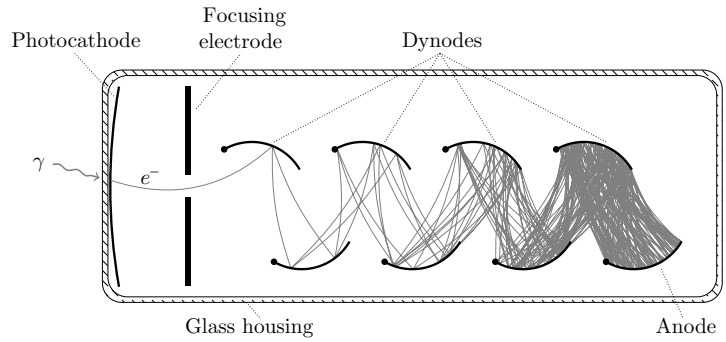


Fig. 1.11 Schematic diagram of a photomultiplier tube. The incident photon is represented by the grey wavy line and electrons by the grey curves.

depend on the energy of the particle (for instance, inhomogeneity of the calorimeter). Typical energy resolutions are $(1-5\%)/\sqrt{E[\text{GeV}]}$ and $(5-20\%)/\sqrt{E[\text{GeV}]}$, respectively, for homogeneous and sampling electromagnetic calorimeters (Fabjan and Gianotti, 2003).

Detection of Visible and Ultraviolet Photons: Photomultipliers

Photomultipliers convert light into a measurable electric current. They are used in many detectors to convert scintillation or Cherenkov light, for instance. Figure 1.11 shows a schematic diagram of a photomultiplier tube (denoted PMT).

When a photon impinges upon a photocathode, an electron, called a *photoelectron*, is extracted by the photoelectric effect. It is focused by an electric guiding field onto the first dynode, where it transfers some of its energy to the electrons in the dynode, producing secondary electrons, which are in turn accelerated toward the next dynode. The process continues by creating an 'avalanche' of electrons, eventually collected by the anode. The voltage between the photocathode and the anode is subdivided by a chain of resistors such that it increases gradually from one dynode to the next. Voltages of the order of 1 000–2 000 volts are typically used to accelerate electrons within the chain of dynodes, with the most negative voltage connected to the cathode and the most positive voltage connected to the anode. The current collected at the anode is then proportional to the number of incident photons. Typically three to five electrons are emitted by a dynode, for each incident electron. Most of the photomultipliers contain 10–14 stages of dynodes. Hence, gains ranging from 10^6 to 10^8 are achievable, allowing the detection of a single photon. A photomultiplier, consisting of N dynodes and producing, on average, δ secondary electrons per dynode, generates a total charge of $\delta^N e$ Coulombs per photoelectron. Because electrons follow different paths, there is a natural variation in their transit time (denoted TTS) up to the anode. Very large photomultipliers can have TTS $\sim$ 5 ns. It thus takes about 5 ns to collect the total charge by the anode. Hence, such a photomultiplier, with $N = 12$ and $\delta = 4$ (these are typical numbers), would produce a current of about 0.5 mA per photoelectron, an amply measurable current.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

Photomultipliers are characterised by their TTS but also by their quantum efficiency, i.e. the number of photoelectrons released by the photocathode divided by the number of incident photons. Depending on the material of the photocathode, the photomultiplier can be sensitive to photons in the infrared to ultraviolet spectrum. Most of the photocathodes are made of semiconductor materials because of their good quantum efficiency, ranging typically from 10% to 30%.

In recent years, silicon photomultipliers (SiPM)[17] have become very popular in particle physics detectors (in Simon, 2019, many recent applications are given). A SiPM is a set of pixels where each pixel (whose typical size is a few tens of micrometres and a SiPM can have over a thousand pixels per mm^2) is a silicon photodiode. However, all pixels are built on a common silicon substrate, and the photodiodes are configured in Geiger mode (i.e., producing a large amplitude in response to the detection of even a single photon), by means of a very strong bias voltage above the breakdown voltage, $V_{\text{bkd.}}$. Instead of producing an avalanche (i.e., a charge carrier multiplication) that leads to a signal proportional to the optical input signal as in the APD, the avalanche is self-sustaining, producing a signal pulse that is not dependent on the intensity of the optical input signal. It is, therefore, an all-or-nothing mode. Once there is an avalanche, the bias voltage automatically reduces to below $V_{\text{bkd.}}$, resulting in the quenching of the avalanche, and quickly returns above that value after a recovery time, in order to be ready to sense new optical input signals [see Buzhana et al. (2003) for more technical details]. Hence, if a SiPM pixel operates as a binary device, the SiPM as a whole is an analogue detector, which can measure the light intensity by counting the numbers of illuminated pixels, and, if the recovery time of each pixel is short enough compared to the duration of the light source (recovery times of the order of a few nanoseconds are possible), by counting the number of times a given pixel is fired. There are many advantages of SiPMs for detectors; in particular, their small size greatly improves the detector granularity compared with standard photomultipliers, while still allowing a single photon to be counted. In addition, they are very fast and insensitive to magnetic fields. Their dynamic range is, however, intrinsically limited by the number of pixels.

Detection of Hadrons: Hadronic Calorimeters

Charged or neutral hadrons with high energy (kinetic energy above 100 MeV) undergo inelastic nuclear reactions with matter that produce secondary hadrons, which in turn generate nuclear reactions and so on. This process produces a *hadronic shower*. The quantity that characterises the development of a hadronic shower is the inelastic *nuclear interaction length* λ_{I} that is analogous to the electromagnetic radiation length, X_0, for electromagnetic showers. It governs the absorption of hadrons by the matter:

$$N(x) = N_0\, e^{-x/\lambda_{\text{I}}},$$

where $N(x)$ is the number of hadrons surviving after passing through a length x in the matter. Generally, λ_{I} is much larger than X_0. For instance, in iron $\lambda_{\text{I}} = 16.8$ cm, whereas

[17] SiPM are also called Multi-Pixel Photon Counter and Geiger-mode avalanche photo diode (MPPC and GAPD), and by many other names depending on the manufacturer.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

$X_0 = 1.76$ cm. Hence, a hadronic shower tends to develop over a larger length scale than an electromagnetic one, and hadronic calorimeters must then be thicker than electromagnetic calorimeters to contain the shower. In practice, the size of hadronic calorimeters with respect to that of electromagnetic ones cannot scale as λ_I/X_0. (If the electromagnetic calorimeter is 1 m thick, the hadronic one should be almost 10 m thick!) Consequently, only a fraction of the longitudinal development of the shower may be contained, with the thickness of hadronic calorimeters being typically of the order of 5–10 interaction lengths (while that of electromagnetic calorimeters ranges from 20 to 30 X_0). Because of large transverse momentum transfers in nuclear interactions, the lateral extension of hadronic showers is also larger than that of electromagnetic showers. Figure 1.10 illustrates the difference between a shower initiated by a photon and a proton in the air (for non-dense media such as air, λ_I and X_0 can be of the same order of magnitude, explaining why the two showers have a similar longitudinal extension in the figure). The dynamics of hadronic showers are complicated. First, about 40% of the energy of the incident energy is dissipated and does not contribute to the production of secondary hadrons (Grupen and Shwartz, 2011). It can go to binding energy, can be transferred to recoils of nuclei or can simply escape detection through the production of neutrinos. Second, nuclear reactions produce many pions, of which about one-third are π^0 that decay into two photons. They induce an electromagnetic shower, mainly concentrated around the hadronic shower core since electromagnetic showers are narrower than hadronic ones. The π^0 production, however, is subject to large fluctuations. Hence, from one event to another, the shower can contain a large or negligible fraction of electromagnetic contribution. Since, in the end, the measured energy is coming from low- energy charged particles (electrons, pions, protons, etc.) interacting with the active material, it is then difficult to optimise a calorimeter to respond similarly to electrons and hadrons. All these complications imply that the energy resolution of hadronic calorimeters is worse than that of pure electromagnetic ones, typically,

$$\frac{\sigma(E)}{E} \sim \frac{(30\% \text{ to } 120\%)}{\sqrt{E[\text{GeV}]}}.$$

The best resolution is generally obtained with calorimeters that use uranium as an absorber, because in ^{238}U, many neutrons and energetic photons are produced by its fission that can eventually generate a visible signal.

Most of hadronic calorimeters are sampling calorimeters. The nuclear interaction length of the active material that produces the visible signal is so large that it cannot contain a significant fraction of the shower. Therefore, layers of absorber material are required. Iron or copper are very common absorbers, while the active material is often composed of plastic scintillator tiles.

Detection of Muons

Muons mainly lose their energy by ionisation. To give an order of magnitude, in copper ($\rho = 8.96$ g/cm^3), an MIP loses about $1.4 \times 8.96 \sim 12.5$ MeV/cm (Fig. 1.4). Hence, to first approximation, a 12.5 GeV muon would travel over 10 m (although the actual range is shorter because muons at that energy are not MIPs, and therefore they lose slightly more energy). Muons are thus highly penetrating particles – a unique feature among charged

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

elementary particles. In other words, they are the only surviving charged particles after travelling a few metres of matter.

Particle detectors at accelerators use this property extensively (in addition to possible identification from the $\mathrm{d}E/\,\mathrm{d}x$ or the Cherenkov effect). In the outermost parts of the detector, beyond the absorbers of the calorimeters, tracking detectors are installed to detect the passage of muons. Since the flux of particles in these regions is strongly reduced compared to the region close to the collision point, gaseous detectors are often used. In those detectors, the muon ionises the atoms of the gas, creating electron–ion pairs. The average energy needed to produce a pair is about 30 eV, with a weak dependence on the gas (compared to the 3 eV in semiconductor devices for electron–hole pairs). A system of anodes and cathodes with a large potential difference amplifies the number of primary electron–ion pairs, creating an avalanche proportional to the initial signal. Generally, a noble gas, such as argon, is chosen since it requires a lower electric field to produce the avalanche. Many muon detectors work on this principle: *cathode strip chambers*, *drift tubes*, *resistive plate chambers*, *gas electron multipliers*, etc. They mostly differ by the structure of their system of anodes or cathodes and the shapes of the latter (wires, strips, plates, etc.), leading to different performance in terms of timing response and accuracy of the position measurement.

Particle detectors for non-accelerator physics generally use other techniques to identify muons. For instance, in water Cherenkov detectors, such as Super-Kamiokande in Japan, muons produce sharp Cherenkov rings (resulting from the projection of the Cherenkov light-cone onto photomultipliers), whereas rings originating from electrons are fuzzier because of the multiple scattering that affects electrons. The analysis of the rings thus makes it possible to identify the particles.

Notice that for very high energy muons (several hundreds of GeV), the main source of energy loss is not ionisation but radiative processes (e^+e^- pair production, bremsstrahlung, and photonuclear interactions).[18] Hence, they can also produce electromagnetic showers (due to e^+e^- pair production and bremsstrahlung) or hadronic showers (due to photonuclear interaction) in calorimeters.

Detection of Neutrinos

Neutrinos are stable particles and light enough to travel over large distances at (almost) the speed of light, even at low energy (cf., Fig. 1.3). However, since they only experience the weak interaction, the probability that they interact with matter turns out to be ridiculously weak. Despite this, there are specific detectors dedicated to their detection, as mentioned in Chapter 9. These detectors usually exploit the production of a charged lepton, coming from the interaction of a neutrino or anti-neutrino with a nucleon. To give an order of magnitude, the cross section (whose exact definition is given in Section 3.5) of the reaction $\nu_e + n \to e^- + p$ is about 0.9×10^{-38} cm^2, for a neutrino energy of 1 GeV (Formaggio and Zeller, 2012). As a typical example, given that the flux of atmospheric neutrinos at that energy is

[18] Photonuclear production of muons is the process $\mu^\pm + \text{nucleon} \to \mu^\pm + \text{hadrons}$, where a virtual photon is exchanged between the muon and the nucleon.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

about 1 $\mathrm{cm}^{-2}\ \mathrm{s}^{-1}$, we obtain an interaction rate per nucleon of $0.9 \times 10^{-38}\ \mathrm{s}^{-1}$. Hence, only detectors containing a huge amount of matter can expect to observe these neutrinos via the detection of the electron or the proton. Since 1 g of matter contains about $N_A = 6 \times 10^{23}$ nucleons (the Avogadro number), a one kiloton detector running one year (3.15×10^7 s) would detect at most 170 neutrinos. This is why modern neutrino detectors, aiming to observing the atmospheric flux, are usually tens of kilotons. The situation is even worse for lower energy neutrinos. At about 1 MeV, the cross section of the neutrino/antineutrino-nucleon inelastic scattering is about $\sigma = 10^{-43}\ \mathrm{cm}^2$ (Giunti and Kim, 2007). The mean free path (introduced in Section 3.5), i.e. the average distance between two interactions, is given by

$$\lambda = \frac{1}{\rho\sigma},$$

where ρ is the density of nucleons. For example, in lead, $\rho = N_A \times 11.3$ (density) $\simeq 7 \times 10^{24}$ nucleons/cm^3, and hence, $\lambda \sim 1.5 \times 10^{18}$ cm. More than one lightyear in lead without an interaction!

Detectors operating at colliders cannot afford to have so much material dedicated to the detection of neutrinos. Therefore, neutrinos usually simply escape those detectors without leaving signals of their presence. Nevertheless, in the centre-of-mass frame of the two incident particles from the collision of the beams, the total momentum should be zero by definition. Hence, if all particle momenta can be observed except those of neutrinos, the missing momentum can be attributed to the presence of neutrinos (assuming, of course, that no other unknown particles escape the detector), i.e.

$$\boldsymbol{p}_\nu \simeq \boldsymbol{p}_{\mathrm{miss}} = -\sum_i \boldsymbol{p}_i,$$

where i runs over the momenta of the observed particles. However, this method gives only a crude estimate of the neutrino momentum, limited by the momentum resolution of the measured particles and the hermiticity of detectors (i.e., their coverage in solid angle, ideally, 4π).

Example of a General-Purpose Detector at a Collider

General-purpose detectors aim to cover the full diversity of reactions issued from beam collisions. They must be able to identify all long-lived particles reaching the detector after a collision (cf. list 1.10). Detectors at accelerators consist of a succession of concentric sub-detectors surrounding the collision point, each specialised in a given task: track determination, momentum measurement, energy measurement, and particle identification. Figure 1.12 shows the layout of a typical detector in the transverse view, i.e. in the view perpendicular to the beam direction. The concentric sub-detectors form the central part of the detector called the *barrel*. The barrel is completed by two *end-caps* closing the barrel from each side to detect particles emitted from all angles to the beam axis (end-caps are not represented in Fig. 1.12). Such a detector is referred to as *hermetic*. It covers nearly a solid angle of 4π around the collision point, the only particles not crossing active elements of the detector being those emitted into the beam pipe (with angles of less than about 1°

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

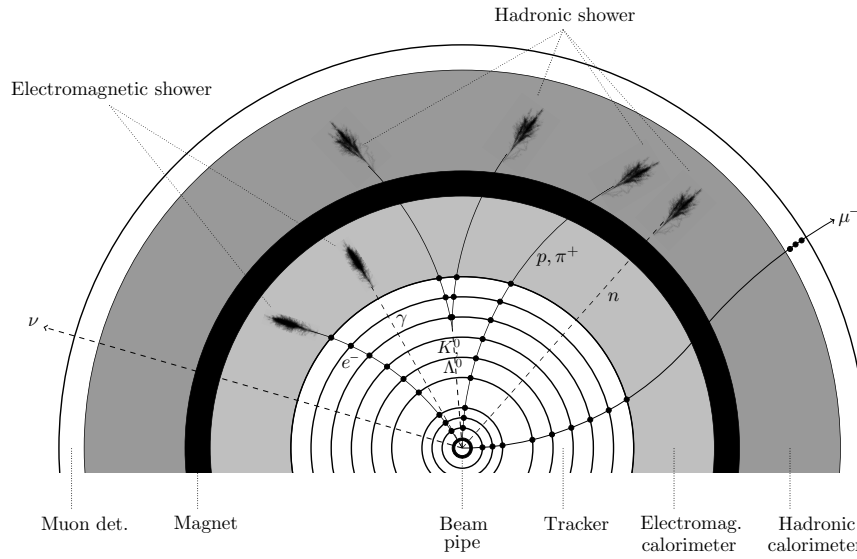


Fig. 1.12 Layout (transverse view) of a general-purpose detector at colliders. The beams circulate in the dimension perpendicular to the page and collide in the centre of the detector. The position of the magnet (a solenoid) in between the two calorimeters is purely indicative; in some detectors, it surrounds the tracker, while in others, it is outside the two calorimeters. The magnetic field is perpendicular to the page, towards the reader. Charged particle trajectories are represented by solid curves, while those of neutral particles are shown by dashed lines.

to the beam axis). Recent examples of such detectors are the ATLAS and CMS detectors operating at the LHC at CERN.

At the very core of the detector, a tracker is in charge of the momentum measurement of charged particles. The tracker determines the curvature of the trajectories of the charged particles in a powerful magnetic field. The innermost layers of the sensitive material (very often silicon pixels) are located radially at a few centimetres from the interaction point, which is, to a first approximation, in the centre of the detector in transverse view. They have the finest granularity of all sub-detectors in order to cope with the high density of particles, and they are crucial to the reconstruction of tracks from short-lived particles. By interpolating tracks towards the centre of the detector, the actual position of the interaction point, usually called the *primary vertex*, can be estimated. Particles such as B hadrons containing a b-quark, which have a significant lifetime of the order of 10^{-12} s, cannot reach those first layers; however, they travel over a few millimetres before decaying. The interpolation of the charged tracks of their decay products allows the identification of a secondary vertex that is displaced with respect to the primary vertex. The distance between the secondary and primary vertices thus gives a measurement of the distance travelled by the meson and indirectly its lifetime.

The next sub-detector encountered by long-lived particles is the electromagnetic calorimeter. It stops electrons or photons by absorbing (ideally all) their energy through the development of electromagnetic showers. An electron is thus distinguished from a photon by the presence of a track in the tracker.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

Depending on the experiment, a solenoid providing a powerful magnetic field along the beam axis (typically between 1 and 4 T) can be found surrounding the tracker, the electromagnetic calorimeter (as shown in Fig. 1.12) or even the hadronic calorimeter (in CMS for instance). The solenoid is often surrounded by steel or iron structures, the 'return yoke', which confines the high magnetic field to the volume of the detector (it provides a flux return path for the central solenoid field). Hence, outside the solenoid, there is a residual magnetic field whose direction is reversed, explaining the strange trajectory of the muon track in Fig. 1.12. The return yoke structures are not represented in Fig. 1.12. They are often inserted in the muon sub-detectors or the calorimeters.

Hadrons reaching the hadronic calorimeter develop a hadronic shower (see Section 'Detection of Hadrons: Hadronic Calormeters'). They are mostly pions, protons and neutrons. A neutron does not leave a track in the tracker, while charged hadrons do. Very long-lived hadrons, such as the K^0 and Λ^0, can reach the tracker when sufficiently boosted, but more likely they decay in the tracker after travelling a significant distance (several tens of centimetres). Since they often decay into two charged hadrons, their signature in the detector is of V-shape, with the vertex clearly separated from the primary vertex (see Fig. 1.12). Incidently, in the early days of particle physics (mid-twentieth century), those particles were called V-particles.

Finally, the outermost sub-detector is the muon sub-detector. Usually, muons are the only charged particles that can pass through the rest of the detector without being stopped (see Section 'Detection of Muons'). Note, however, that charged pions produced in a hadronic shower in the hadronic calorimeter can sometimes leak into the muon detector, mimicking a muon signal.

1.5 What Is a Measurement?

Physics is an experimental science. Even the most beautiful theories must be falsifiable (to claim to be science) and hence should be confronted, at least conceptually, with experimental measurements. Experimental measurements can be used to estimate a quantity from some data, and the question then arises of the best estimate of that quantity and its uncertainty. Measurements can also be used to test a hypothesis that is based on a particular model and to test the consistency of the data with that model. This section assumes that the reader already has an elementary knowledge of probability theory and statistics. Only a few topics relevant to particle physics are covered in this section, without proving them in depth. The reader can (and should) consult statistics textbooks, such as Barlow (1989), James (2006), Bohm and Zech (2010) and Lyons (1986) to fill in the gaps.

1.5.1 Generalities

Randomness is inherent in particle physics because it is built upon quantum mechanics. For instance, a muon decays *on average* in 2.2 microseconds, but sometimes it takes less and sometimes more. We can calculate the probability that it decays at a given time, but

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

we cannot say precisely when it will happen. Even the signal collected by a detector has a degree of randomness: a charged particle passing through a medium has a certain probability of ionising the atoms of that medium, which induces intrinsic fluctuations in the collected current.

Continuous random processes are described by *probability density functions* (p.d.f.). If x is a random variable and $p(x)$ is its p.d.f. normalised to unity, then $p(x)\ \mathrm{d}x$ is the probability of finding x between x and $x + \mathrm{d}x$. The first two moments of the distribution, the *mean* μ and *variance* σ^2, are given by

$$\mu = E[x] = \int xp(x)\ \mathrm{d}x, \tag{1.35}$$

where $E[x]$ is the expectation value and

$$\sigma^2 = E[(x-\mu)^2] = E[x^2] - \mu^2. \tag{1.36}$$

The square root of the variance, σ, is the *standard deviation*, and it represents a measure of the width of the distribution. For outcomes of a process depending on several random variables, let us say x and y, the *covariance* is

$$\mathrm{cov}[x,y] = E[xy] - E[x]E[y] = E[(x-\mu_x)(y-\mu_y)], \tag{1.37}$$

where μ_x and μ_y are the means of x and y, respectively. These mean values are given by Eq. (1.35), inserting a multivariable p.d.f., $p(x,y)$, and performing a double integral over x and y. The variance of x is trivially deduced from the covariance, with $\sigma_x^2 = \mathrm{cov}[x,x]$. It is convenient to introduce the dimensionless *correlation coefficient*

$$\rho_{xy} = \frac{\mathrm{cov}[x,y]}{\sigma_x \sigma_y}, \tag{1.38}$$

which varies between -1 and $+1$. When x and y are independent, namely when the p.d.f. is factorisable, i.e. $p(x,y) = p_x(x)p_y(y)$, the covariance and correlation coefficient are zero.[19] It is standard to have random variables that depend on other random variables. If x and y are two random variables, and $x'(x,y)$ and $y'(x,y)$ are two variables depending on x and y, the covariance of x' and y' is related to that of x and y via the formula

$$\begin{aligned}\mathrm{cov}[x',y'] = \ & \frac{\partial x'}{\partial x}\frac{\partial y'}{\partial x}\mathrm{cov}[x,x] + \frac{\partial x'}{\partial y}\frac{\partial y'}{\partial y}\mathrm{cov}[y,y] \\ & + \frac{\partial x'}{\partial x}\frac{\partial y'}{\partial y}\mathrm{cov}[x,y] + \frac{\partial x'}{\partial y}\frac{\partial y'}{\partial x}\mathrm{cov}[y,x],\end{aligned} \tag{1.39}$$

where all derivatives are evaluated with $x = \mu_x, y = \mu_y$. The variance, $\sigma_{x'}^2 = \mathrm{cov}[x',x']$, is then

$$\sigma_{x'}^2 = \left(\frac{\partial x'}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial x'}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial x'}{\partial x}\frac{\partial x'}{\partial y}\rho_{xy}\sigma_x\sigma_y. \tag{1.40}$$

The extension of Formulas (1.39) and (1.40) to more variables, $x_1, \ldots, x_n$, is straightforward using the covariance matrix

$$V_{ij} = \mathrm{cov}[x_i, x_j] = \rho_{ij}\sigma_i\sigma_j. \tag{1.41}$$

[19] The converse is not necessarily true: the covariance can be 0 when x and y depend on each other. This is true, for example, if x is uniformly distributed between 0 and 1 and $y = \pm x$, where the sign is randomly chosen.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

For instance, Eq. (1.40) becomes

$$\sigma_{x_i'}^2 = \sum_{k,l} \frac{\partial x_i'}{\partial x_k} \frac{\partial x_i'}{\partial x_l} V_{kl} = \sum_k \left(\frac{\partial x_i'}{\partial x_k} \right)^2 \sigma_k^2 + 2 \sum_{k>l} \frac{\partial x_i'}{\partial x_k} \frac{\partial x_i'}{\partial x_l} \rho_{kl} \sigma_k \sigma_l. \tag{1.42}$$

Notice that these formulas are an approximation when the variables x' and y' are non-linear combinations of x and y.

Among the infinite possibilities of p.d.f., the Gaussian (or normal) distribution and the Poisson distribution are the most often encountered. The Gaussian distribution is of the utmost importance because of the *central limit theorem*, which states that the sum of n random variables distributed according to any p.d.f. with finite mean and variance has a p.d.f. approaching the Gaussian distribution for large n. As many measurements are based on averaging samples of data, Gaussian distributions are often observed. The well-known formula of the Gaussian density function with mean μ and variance σ^2 (and thus standard deviation, σ) is

$$G(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right]. \tag{1.43}$$

The Gaussian density function is represented in Fig. 1.13. The probability that the random variable x lies in the interval $[a, b]$ is given by

$$p_{a,b} = \int_a^b G(x;\mu,\sigma)\ \mathrm{d}x = F_G(b) - F_G(a), \tag{1.44}$$

where $F_G(x) = \int_{-\infty}^{x} G(x';\mu,\sigma)\ \mathrm{d}x'$ is the *cumulative function* of the Gaussian. Conversely, the probability of being outside this interval is $1 - p_{a,b}$. The canonical values in terms of numbers of standard deviations from the mean value, $\mu \pm n\sigma$, are given in Fig. 1.13. With the change of variable $x \to t = (x-\mu)/(\sqrt{2}\sigma)$, $p_{a,b}$, with $a = \mu - n\sigma$ and $b = \mu + n\sigma$, takes the expression $p_{a,b} = \mathrm{erf}(n/\sqrt{2})$, where $\mathrm{erf}(x)$ is the error function $\mathrm{erf}(x) = 1/\sqrt{\pi} \times \int_0^x \exp(-t^2)\ \mathrm{d}t$.

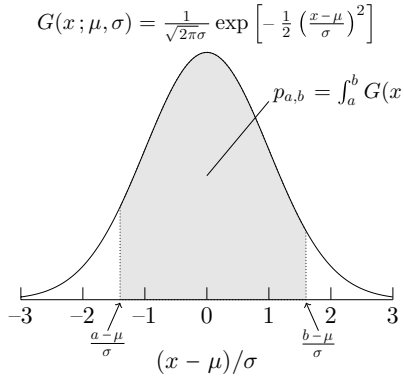
Whereas the central limit theorem states that a variable that is the *sum* of a large number of random variables is described by the Gaussian (or normal) density function, a variable that is the *product* of random variables is described by the log-normal density function. The logarithm of such a variable would follow a Gaussian distribution $G(\ln(x);\mu,\sigma)$, and thus the p.d.f. of x satisfies $f(x;\mu,\sigma)\ \mathrm{d}x = G(\ln(x);\mu,\sigma)\ \mathrm{d}(\ln(x))$.[20] It follows that the log-normal density function is

$$f(x;\mu,\sigma) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \left(\frac{\ln(x)-\mu}{\sigma} \right)^2 \right], \tag{1.45}$$

with $x > 0$. The mean and standard deviation are $e^{\mu+\sigma^2/2}$ and $e^{\mu+\sigma^2/2}\sqrt{e^{\sigma^2}-1}$, respectively. The log-normal distribution is encountered in processes with many small multiplicative variations. A typical example is shower development. An electromagnetic shower

[20] In general, if $\boldsymbol{x}$ is a vector of random variables described by the p.d.f. $p_x(\boldsymbol{x})$ and $\boldsymbol{y}$ is a vector of random variables obtained by the one-to-one correspondence $\boldsymbol{y} = f(\boldsymbol{x})$, then the p.d.f. of $\boldsymbol{y}$ is $p_y(\boldsymbol{y}) = p_x(\boldsymbol{x})|J| = p_x(f^{-1}(\boldsymbol{y}))|J|$, where $|J|$ is the determinant of the Jacobian transformation, i.e. $J_{ij} = \partial x_i/\partial y_j$. In the case of a single random variable, the relation is equivalent to $p_y(y)\ \mathrm{d}y = p_x(x)\ \mathrm{d}x$.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press



Interval a	Interval b	$p_{a,b}$	$1-p_{a,b}$
$\mu-\sigma$	$\mu+\sigma$	68.27%	31.73%
$\mu-2\sigma$	$\mu+2\sigma$	95.44%	4.55%
$\mu-3\sigma$	$\mu+3\sigma$	99.73%	2.7×10^{-3}
$\mu-5\sigma$	$\mu+5\sigma$	99.99%	5.7×10^{-7}
$\mu-1.64\sigma$	$\mu+1.64\sigma$	90%	10%
$\mu-1.96\sigma$	$\mu+1.96\sigma$	95%	5%
$-\infty$	$\mu+1.28\sigma$	90%	10%
$-\infty$	$\mu+1.64\sigma$	95%	5%

Fig. 1.13 The Gaussian density function and the fraction of values, with $p_{a,b}$, lying within an interval $[a, b]$.

results mainly from the multiplication of the pair-production and bremsstrahlung processes (see Section 'Detection of High-Energy Electrons and Photons: Electromagnetic Calorimeters'), each with their corresponding energy loss. Hence, fluctuations in shower size at a given depth, in a sample of showers of the same energy, are approximatively log normal (Gaisser, 1990, chapter 15).

The Poisson density function is also very often encountered in physics experiments. It describes the probability of n events occurring (n, an integer number) given that the mean expected number is μ (a real positive number). The Poisson p.d.f.[21] is

$$P(n;\mu) = \frac{\mu^n}{n!}e^{-\mu}. \tag{1.46}$$

Each event is understood to occur independently of one another. It can be shown that the Poisson distribution is a limiting case of the binomial distribution.[22] By construction, not only the mean of the Poisson distribution is given by μ but also its variance (the standard deviation is thus $\sqrt{\mu}$). The Poisson distribution then depends on a single number. Although the Poisson distribution is discrete, it is normalised to unity in the sense that $\sum_{n=0}^{\infty} P(n;\mu) = 1$. A frequent application is when the mean rate of a process (mean number of reactions per second) is known, for example, the decay rate of a particle, but one wishes to evaluate the probability of observing n events during the time t. If the rate is λ, then the mean number of events during t is $\mu = \lambda t$, and the probability of observing n events is thus given by $P(n;\lambda t)$. An even more frequent application is when data are sampled, i.e. when data are analysed in discrete intervals (called bins), for example, bins in the energy channel of a detector, bins in time of arrival of a particle, etc. One then counts the number of events in each bin. In most of these cases, this number is assumed to follow

[21] It is formally not a p.d.f. since the Poisson distribution is not continuous. For discrete random variables, it is called a probability mass function (p.m.f.), and it directly gives a probability.

[22] Probably, the binomial distribution is already well known by the reader. It describes cases where there are only two possible outcomes (such as heads and tails in a coin toss). If one calls one of the two outcomes 'success', then the probability of obtaining r success in N independent trials is $\frac{N!}{r!(N-r)!}p^r(1-p)^{n-r}$, where p is the intrinsic probability of success. The Poisson distribution is obtained when $p \to 0$, $N \to \infty$ and $N \times p = \mu$.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

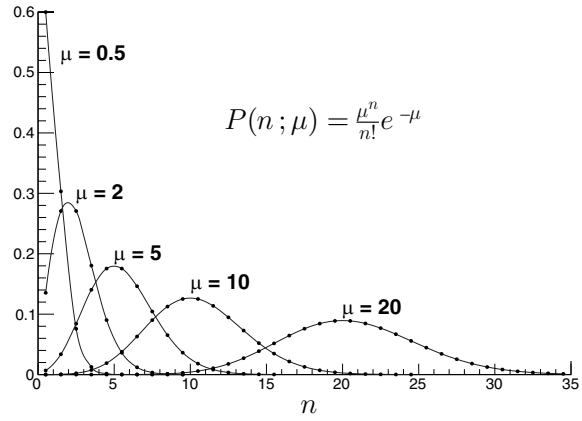


Fig. 1.14 The Poisson distribution. The discrete points for each value of μ are connected by smooth curves for clarity.

a Poisson distribution (counting can never be negative). If n_i events are observed in bin i, assuming that n_i represents the mean number of events of the Poisson distribution of that bin, the uncertainty is then $\sqrt{n_i}$, according to the standard deviation of the Poisson distribution. Notice that it is the estimated uncertainty on the mean of the underlying Poisson distribution. The quantity $\sqrt{n_i}$ is therefore not the error on the number of observed events, which is assumed to be an error-free count. Figure 1.14 presents the Poisson distribution for various values of μ. Notice that the distribution is not symmetric around μ. However, it becomes more symmetric as μ increases. In fact, when μ is large (in practice, above 20), the Poisson distribution $P(n;\mu)$ approaches the Gaussian distribution $G(n;\mu, \sigma = \sqrt{\mu})$.

1.5.2 Measurement Uncertainties

Measurement uncertainties (or errors, both words are used equivalently) are classified into two categories: *statistical uncertainties* often called random uncertainties and *systematic uncertainties* (in short systematics). The former arise, for example, from the inherent statistical nature of the phenomena studied in particle physics. The limiting ability of any device to give measurement with an infinite accuracy also contributes to these kinds of errors. In general, statistical uncertainties are supposed to follow a known statistical distribution. They all share the characteristics that sequential measurements are statistically uncorrelated. The precision is thus improved by combining several measurements. Indeed, if r_i is the result of the ith measurement and σ_{r_i}, its uncertainty, the error resulting from the average of N measurements is, according to Eq. (1.42),

$$\sigma^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sigma_{r_i}^2.$$

Assigning to all measurements the same error σ_r, it follows that $\sigma = \sigma_r/\sqrt{N}$, scaling with the inverse of the square root of the number of measurements. Notice that, in general, when

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

each measurement has a different uncertainty, it is more appropriate to perform a weighted average

$$r = \frac{1}{\sum_i w_i} \sum_{i=1}^{N} w_i r_i,$$

where $w_i = 1/\sigma_{r_i}^2$ (justified in Problem 1.7). Assuming no correlation in Eq. (1.42), the uncertainty of r is then

$$\sigma^2 = \frac{1}{\left(\sum_i w_i\right)^2} \sum_{i=1}^{N} w_i^2 \sigma_{r_i}^2 = \frac{1}{\sum_i w_i}. \tag{1.47}$$

Unlike statistical errors, systematic uncertainties do not necessarily improve with more data. They generally represent a possible bias, mistake, etc. They are not directly due to the statistic of the data. A simple example is a thermometer, whose 0° is actually shifted by a constant value. All measurements will be affected by that value. A more realistic example in the context of particle physics is a calibration error of a calorimeter that affects the energy scale, i.e. the response of the calorimeter to the energy of the incident particle. Another example is a measurement using inputs from a theoretical model that is actually wrong or inputs from a simulation that does not describe properly the real data. Systematic errors affect different measurements made in identical conditions in the same manner. Hence, repeated measurements do not reduce systematic errors. It is always delicate to estimate systematic errors. There is no universal recipe, and it often relies on the experience of the experimentalist who can guess what parameter has a significant effect on the final result of an analysis. A case often encountered is an external input parameter with known uncertainty σ. For example, the luminosity of an accelerator. The resulting systematic uncertainty on the analysis is then obtained by varying the input parameter by $\pm\sigma$. The deviation from the initial analysis result is then considered as a systematic uncertainty for this parameter.

At the end of the analysis, all sources of systematic uncertainties have to be combined with the statistical error. Implicitly we assume that the value x that is measured results from the summation, $x = x_{\text{true}} + \Delta x_{\text{stat}} + \sum_i \Delta x_{i,\text{syst}}$, where Δx_{stat} and $\Delta x_{i,\text{syst}}$ have the respective errors σ_{stat} and $\sigma_{i,\text{syst}}$. No correlation is assumed between the statistical uncertainty and the sources of systematic uncertainties. Formula (1.42) then yields

$$\sigma_x^2 = \sigma_{\text{stat}}^2 + \underbrace{\sum_{k,l} \rho_{kl}\, \sigma_{k,\text{syst}} \sigma_{l,\text{syst}}}_{\equiv \sigma_{\text{syst}}^2}.$$

The final result is then generally reported as $r = x \pm \sigma_{\text{stat}} \pm \sigma_{\text{syst}}$. For example, when the Higgs boson was discovered by the ATLAS and CMS experiments in 2012, the publication CMS Collaboration (2012b) reported the mass of the boson as $m_H = 125.3 \pm 0.4$ (stat.) $\pm$ 0.5 (syst.) GeV/c^2. Separately quoting the statistical and systematic uncertainties has the advantage of indicating whether taking more data would significantly reduce the global uncertainty.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

1.5.3 Parameter Estimation: Maximum Likelihood and Least-Squares Methods

As physicists, we very often measure quantities that we want to compare with a model. The model may depend on parameters that we would like to evaluate. Two general methods are briefly described in this section: the maximum likelihood and least-squares methods.

The Maximum Likelihood Method

Let $\boldsymbol{x} = x_1, \ldots, x_n$ be n independent observations. We assume that they all follow the same p.d.f., $f(x;\theta)$, which depends on a parameter θ that we wish to evaluate. The joint probability for obtaining $\boldsymbol{x}$ is the product of the p.d.f. (there should be an extra $n!$ if the order of observations does not matter, but we will see below that this factor will not play any role). When this joint probability is interpreted as a function of the parameter of interest, it is called the *likelihood function*,

$$L(\theta) \equiv L(\boldsymbol{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{1.48}$$

Notice that $L(\theta)$ is not the p.d.f. of θ. Let us call $\hat{\theta}$ an estimator of θ. One can show that when n becomes infinite, then $\hat{\theta}$, given by the global maximum of the likelihood, is a *consistent* estimator, i.e. it converges to the true value of θ. It is *unbiased*, i.e. its expectation value is θ, and its variance converges to the minimum possible variance for an unbiased estimator (it is then qualified as *efficient*). Beware that these nice properties are only true in the asymptotic limit, i.e. when n becomes infinite. For small n, the estimator is usually biased. The name of the method, the maximum likelihood, may suggest that the estimate is the most likely value, whereas it is actually the estimate that makes the data most likely. In practice, it is more convenient to find the global minimum of the negative log-likelihood function (because it converts a product into a sum, which is easier to handle in terms of numerical accuracy),

$$\mathcal{L}(\theta) = -\ln\left(L(\theta)\right) = -\sum_{i=1}^{n} \ln\left(f(x_i;\theta)\right). \tag{1.49}$$

The estimator $\hat{\theta}$ is then a solution of

$$\left.\frac{\partial}{\partial\theta}\mathcal{L}(\theta)\right|_{\theta=\hat{\theta}} = -\sum_{i=1}^{n} \left.\frac{\partial}{\partial\theta}\ln\left(f(x_i;\theta)\right)\right|_{\theta=\hat{\theta}} = 0. \tag{1.50}$$

In general, there is no analytic solution to this equation, and the solution must be found numerically. Once $\hat{\theta}$ is found, we can expand $\mathcal{L}(\theta)$ around this value and obtain

$$\mathcal{L}(\theta) = \underbrace{\mathcal{L}(\hat{\theta})}_{\mathcal{L}_{\min}} + \frac{1}{2!}\left.\frac{\partial^2\mathcal{L}(\theta)}{\partial\theta^2}\right|_{\theta=\hat{\theta}} (\theta-\hat{\theta})^2 + O(\theta-\hat{\theta})^3. \tag{1.51}$$

Let us set

$$\sigma_{\hat{\theta}} = \left(\frac{\partial^2\mathcal{L}(\theta)}{\partial\theta^2}\right)^{-1/2}\Bigg|_{\theta=\hat{\theta}}. \tag{1.52}$$

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

Equation (1.51) then becomes

$$\mathcal{L}(\theta) = \mathcal{L}_{\min} + \frac{1}{2}\left(\frac{\theta - \hat{\theta}}{\sigma_{\hat{\theta}}}\right)^2 + O(\theta - \hat{\theta})^3. \tag{1.53}$$

Or equivalently,

$$L(\theta) \simeq L_{\max} \times \exp\left(-\frac{1}{2}\left(\frac{\theta - \hat{\theta}}{\sigma_{\hat{\theta}}}\right)^2\right).$$

The quantity $\sigma_{\hat{\theta}}$ in Eq. (1.52) represents the standard deviation of $\hat{\theta}$ only when $\mathcal{L}$ is reasonably parabolic near $\hat{\theta}$. Equivalently, $\sigma_{\hat{\theta}}$ can be obtained from Eq. (1.53) since for $\theta = \hat{\theta} \pm \sigma_{\hat{\theta}}$,

$$\mathcal{L}(\hat{\theta} \pm \sigma_{\hat{\theta}}) = \mathcal{L}_{\min} + 1/2 \;\text{ or }\; L(\hat{\theta} \pm \sigma_{\hat{\theta}})/L_{\max} = \exp(-1/2). \tag{1.54}$$

The standard deviation is then obtained when $\mathcal{L}$ increases by 1/2 from its minimum value. Similarly, the n standard deviations are obtained when $\mathcal{L}$ increases by $n^2/2$ from its minimum value. Even for non-parabolic forms, Eq. (1.54) is used to derive the error on $\hat{\theta}$. In such a case, the error may not be symmetric around $\hat{\theta}$. We can do so because estimators obtained with the likelihood method are *invariant* under parameter transformations: if $\alpha = g(\theta)$, with g being a one-to-one transformation function, then the estimator of α that maximises the likelihood is just $\hat{\alpha} = g(\hat{\theta})$. Hence, if $\mathcal{L}(\theta)$ is not parabolic around $\hat{\theta}$, we can perform the appropriate transformation $\alpha = g(\theta)$, such that $\mathcal{L}(\alpha)$ becomes parabolic. The invariance property not only applies to the maximum of the likelihood L but also to its relative values (James, 2006) and thus to the differences of its logarithm $\mathcal{L}$. Hence, the values of α satisfying $\mathcal{L}(\alpha) = \mathcal{L}(\hat{\alpha}) + 1/2$ correspond to the values of θ satisfying $\mathcal{L}(\theta) = \mathcal{L}(\hat{\theta}) + 1/2$, and the 1σ domain (or any confidence interval) can be found without explicitly finding the transformation function g. Figure 1.15 illustrates the situation for a non-parabolic shape, where the 68.3% confidence interval $[\theta_-, \theta_+]$ containing the true value of θ (cf. table in Fig. 1.13) is not symmetric around $\hat{\theta}$. The measured value of θ is then noted with its errors as $\hat{\theta}^{+\sigma_+}_{-\sigma_-}$ instead of $\hat{\theta} \pm \sigma_{\hat{\theta}}$. For instance, the mass of the u quark in Particle Data Group (2022) is noted $m_u = 2.16^{+0.49}_{-0.26}$ MeV. When the number of observations is large, $\mathcal{L}$ gets closer to a parabola (a consequence of the central limit theorem), and σ_-, σ_+ in Fig. 1.15 converge to $\sigma_{\hat{\theta}}$, the standard deviation of $\hat{\theta}$ in the parabolic case.

For a set of parameters $\boldsymbol{\theta} = \theta_1, \ldots, \theta_m$, there are m equations such as Eq. (1.50) to solve, i.e.

$$\frac{\partial}{\partial \theta_k}\mathcal{L}(\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\sum_{i=1}^{n} \frac{\partial}{\partial \theta_k} \ln\left(f(x_i;\boldsymbol{\theta})\right)\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0, \quad k = 1, \ldots, m.$$

The standard deviation of the parameters is obtained for the parabolic case by finding the covariance matrix where the elements of its inverse are given by

$$\left(V^{-1}\right)_{ij} = \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{1.55}$$

For non-parabolic cases, Eq. (1.54) extended to the m-dimension space must be applied.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

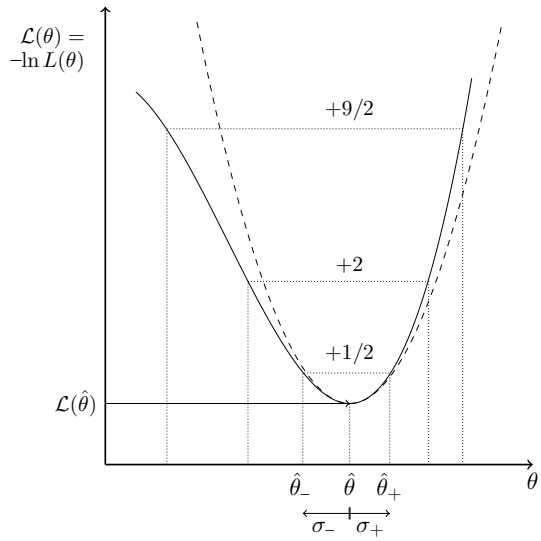


Fig. 1.15 The negative log-likelihood function (solid curve) and the parabolic approximation around the minimum value (dashed line). The asymmetric errors are σ_+ and σ_-, corresponding to the '1σ domain' (68.3%), obtained with $\mathcal{L}(\hat{\theta}) + 1/2$. The '$2\sigma$ and 3σ domains' (95.4% and 99.7%, respectively) are obtained with $\mathcal{L}(\hat{\theta}) + 2$ and $\mathcal{L}(\hat{\theta}) + 4.5$, respectively.

For processes where the number of events is not fixed, if one wants to compare the observed number of events n to a theory prediction ν that might partially depend on the other parameters $\boldsymbol{\theta}$, an additional Poisson term can be incorporated in the likelihood function. Equation (1.48) then becomes

$$L(\nu, \boldsymbol{\theta}) \equiv L(\boldsymbol{x}\,; \nu, \boldsymbol{\theta}) = e^{-\nu} \frac{\nu^n}{n!} \prod_{i=1}^{n} f(x_i\,; \boldsymbol{\theta}). \tag{1.56}$$

This variation of the maximum likelihood method is called the *extended maximum likelihood*. The unknown parameters $\boldsymbol{\theta}$ are now not only encoded in the shape of the data distribution but also in the number of events. The estimate of the expected number of events $\hat{\nu}$ is then obtained by the minimisation of

$$\mathcal{L}(\nu, \hat{\boldsymbol{\theta}}) = -\ln(L(\nu, \hat{\boldsymbol{\theta}})) = \nu - n \ln n - \sum_{i=1}^{n} \ln\left(f(x_i\,; \theta)\right) + \text{constant}$$

with respect to all free parameters. If ν does not depend on $\hat{\boldsymbol{\theta}}$, the constraint $\partial \mathcal{L} / \partial \nu = 0$ yields $\hat{\nu} = n$, and the values $\hat{\boldsymbol{\theta}}$ are the same as those obtained from the standard maximum likelihood method. Otherwise, the additional constraint improves the accuracy on $\hat{\boldsymbol{\theta}}$.

The Least-Squares Method

Let us suppose that we have n observations of two variables, x and y. The x_i variables ($i \in [1, n]$) are assumed to have negligible uncertainties, whereas y_i has an uncertainty

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

σ_i. A situation often encountered is the case of binned data, where x is displayed in a histogram with n bins, which represents the observed distribution of x. For bin i, $x = x_i$ and y_i would be the bin content, i.e. the number of events corresponding to the value x_i. Let us imagine that the relation between x_i and y_i is given by theoretical supposition, $s(x_i;\theta)$, which depends on a parameter θ that we wish to estimate. For example, x represents the set of decay times of many identical particles, and $s(x_i;\theta)$ would be the exponential function, with θ being the lifetime of the particle. If $s(x_i;\theta)$ describes reasonably well the data, the variable y_i should be very close to $s(x_i;\theta)$. The least-squares method consists in minimising the squares of the residuals,

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i - s(x_i;\theta)}{\sigma_i} \right]^2, \tag{1.57}$$

with respect to the parameter of interest, i.e.

$$\left. \frac{\partial \chi^2}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0. \tag{1.58}$$

Formula (1.57) assumes that the y_i variables are uncorrelated. If it is not the case, one has to take into account their correlation, using the correlation matrix V, and thus Eq. (1.57) becomes

$$\chi^2 = \sum_{i,j=1}^{n} [y_i - s(x_i;\theta)] \left(V^{-1}\right)_{ij} [y_j - s(x_j;\theta)]. \tag{1.59}$$

Strictly speaking, χ^2 in Eq. (1.57) or (1.59) is a true *chi-square* only if y_i is Gaussian distributed with mean $s(x_i;\theta)$ and standard deviation σ_i. It is generally a reasonable assumption since we saw previously that many distributions with large statistics converge to a Gaussian distribution. In such a case, χ^2 follows a distribution called the *chi-square distribution* given by the p.d.f.

$$f_\chi(\chi^2;n) = \frac{2^{-\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} \left(\chi^2\right)^{\frac{n-2}{2}} \exp\left(-\frac{\chi^2}{2}\right), \tag{1.60}$$

which has a mean value n and a variance $2n$. The Gamma function is defined as $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, \mathrm{d}t$ and corresponds to $(x-1)!$ when x is an integer. For large n (typically above 30), $f_\chi(\chi^2;n)$ approaches a Gaussian p.d.f. Notice that because of the constraint (1.58), we adjust one parameter, reducing the number of degrees of freedom (n.d.f.) by one unit. Indeed, θ could be chosen to perfectly match one of the y_i, and hence the variability of χ^2 will not be due to n independent random variables but to $n-1$. Hence, if $\chi^2(\theta)$ in Eq. (1.57) follows the chi-square distribution $f_\chi(\chi^2;n)$, $\chi^2(\hat{\theta})$ actually follows $f_\chi(\chi^2;n-1)$. Similarly, if m parameters $\boldsymbol{\theta} = \theta_1, \ldots, \theta_m$ are estimated, the chi-square distribution is given by $f_\chi(\chi^2;n-m)$.

The least-squares method is rather general since knowledge of the actual distribution of y_i is not required, only of its variance. It is thus simple to implement. (An example of its application to estimate parameters is proposed in Problem 1.8.) However, if the distribution of y_i is known, it is better to use the maximum likelihood method since it will be more

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

accurate. Notice that both give the same estimator if y_i is Gaussian distributed. Indeed, in such a case, the p.d.f. of the y_i (uncorrelated) measurements is

$$f(\mathbf{y};\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{y_i - s(x_i;\theta)}{\sigma_i}\right)^2\right],$$

and thus the negative log-likelihood reads

$$\mathcal{L}(\mathbf{y};\theta) = -\ln\left(f(\mathbf{y};\theta)\right) = \chi^2/2 + \sum_{i=0}^{n} \ln\left(\sqrt{2\pi}\sigma_i\right). \tag{1.61}$$

Hence, minimising $\mathcal{L}$ is equivalent to minimising χ^2 (since only χ^2 depends on θ). We can then take advantage of what we have learned with the likelihood to deduce that, due to Eq.(1.54), the standard deviation error of the estimated parameter θ in the least-squares method is obtained when $\chi^2 = \chi^2_{\min} + 1$, while the domain at n-sigma is obtained with $\chi^2 = \chi^2_{\min} + n^2$. Moreover, in the specific case of Gaussian distributed variables, the transposition of Eq. (1.55) into the least-squares language is

$$\left(V^{-1}\right)_{ij} = \frac{1}{2}\left.\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{1.62}$$

and the standard deviation of the parameters can be recovered directly from the reading of the covariance matrix (at the cost of a matrix inversion). Both methods require the minimisation (or maximisation) of quantities, which generally cannot be performed analytically and require numeric tools. In the high energy physics community, a software package called ROOT (Brun and Rademakers, 1996) is extensively used. It provides many tools and has the great advantage to be free and open source.[23]

The maximum likelihood and the least-squares methods are extensively used to estimate parameters with their uncertainties. One may wonder what is the meaning of the parameter uncertainty or its confidence intervals since, in principle, a parameter (even if unknown) is not expected to be a random variable. When probabilities are interpreted from a frequentist point of view,[24] quoting that $\theta = \hat{\theta} \pm \sigma_\theta$ means that out of N measurements of the parameter θ, one should expect that on average about 68.3% contain the true value within their error intervals. For Bayesians, the parameter θ has a p.d.f. and is therefore seen as a random variable for which σ_θ has a clear interpretation.

1.5.4 Model and Hypothesis Testing

Very often, one wants to test hypotheses such as 'Is there an unknown particle in my data?', 'Is this track a muon or an electron?' and 'Does the energy loss increase logarithmically

[23] See https://root.cern.

[24] There are two schools of thought among statisticians: frequentist and Bayesian. For frequentists, if an experiment is reproduced N times in the same condition, then the probability of a given outcome is the number of times it is observed divided by N when N becomes infinite. It is intuitive, but, in practice, N can never be infinite. Moreover, how should one define a probability, when an experiment can only be realised once, as, for example, the big bang? For Bayesians, probabilities are interpreted as a degree of belief that something will happen. It could be measured, for instance, by considering the odds offered for a bet. This approach extends to more conceptual objects, such as the probability of a theory. Hence, there is a degree of subjectivity, but the definition of probabilities does not suffer from frequentist limitations. Both approaches are complementary.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

with the velocity of the particle?'. In the first two examples, we compare two hypotheses: physics beyond the Standard Model against the physics of the Standard Model and track left by a muon against track left by an electron. The question arises as to which hypothesis is more likely, and one should have clear criteria to claim a discovery or refute it. In the third example, there is no unique alternative hypothesis to which we can compare the hypothesis 'the energy loss increases logarithmically with the velocity of the particle'. Actually, there are even an infinite number of alternative hypotheses. This last case is known as the *goodness of fit*. In all cases, one then needs to define criteria, i.e. a test statistic, to determine the level of agreement of a hypothesis with the observation.

Goodness of Fit

When a fit is performed, i.e. the adjustment of parameters $\hat{\boldsymbol{\theta}}$ of a model using a functional form $s(\boldsymbol{x}, \boldsymbol{\theta})$ to the data $\boldsymbol{x}$, the hypothesis we usually want to test is whether the model is consistent with the data. The χ^2-test is the most popular test statistic [but there are others, see the textbooks James (2006) and Barlow (1989), for instance]. It consists in calculating χ^2 with Eq. (1.59) and evaluating the p-value, i.e. the probability to get χ^2 values that are equal to or greater than the actual value observed in the data, χ^2_{obs}. Since χ^2 is supposed to follow the chi-square probability function (1.60), the p-value is defined as

$$p\text{-value} = \int_{\chi^2_{\text{obs}}}^{\infty} f_\chi(\chi^2; n_{\text{dof}})\, \mathrm{d}\chi^2, \tag{1.63}$$

where n_{dof} is the number of degrees of freedom. `ROOT` implements this calculation for us with the function `TMath::Prob(Double_t chi2, Int_t ndf)`. If m parameters $\hat{\boldsymbol{\theta}}$ were previously adjusted with the same n data points $\boldsymbol{x}$, then $n_{\text{dof}} = n - m$. Moreover, if the least-squares method was used, then $\chi^2_{\text{obs}} = \chi^2_{\text{min}}$. The p-value, being built with the data, is itself a random variable. By construction, it lies between 0 and 1 and is uniformly distributed (if the errors are really Gaussian distributed). A large value of χ^2_{obs}, thus of the squares of the residuals, should indicate a poor fit of the model to the data and leads to a small p-value. It could also be due to an underestimation of errors. Conversely, a very low χ^2_{obs} value is probably due to an overestimation of errors. Assuming that the errors are correctly estimated, the model described by the function $s(\boldsymbol{x}, \boldsymbol{\theta})$ can then be rejected if

$$p\text{-value} < \alpha, \tag{1.64}$$

where $1 - \alpha$ is called the *confidence level*. The choice of α is subjective, but the standard values of α are 1%, 5% or 10%. Instead of the p-value, one can use the following rule of thumb as a quick check: since according to the chi-square p.d.f., Eq. (1.60), the mean value of χ^2 is n_{dof} and the standard deviation is $\sqrt{2n_{\text{dof}}}$, one expects χ^2_{obs} to be reasonably close to n_{dof}. More specifically, the rule of thumb consists in checking that $|\chi^2_{\text{obs}} - n_{\text{dof}}|/\sqrt{2n_{\text{dof}}}$ is less than or similar to 1.

When fit parameters are determined with the least-squares method, it is natural to use the χ^2-test to check the goodness of fit. But if the maximum likelihood method is used with unbinned data, how can we check the quality of the description of the data by the model? A possible solution is to bin the data and perform a χ^2-test using the best-fit parameters

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

$\hat{\boldsymbol{\theta}}$ obtained with the first method to predict the bin content. Let us denote by n the number of bins, by n_i the number of events in bin i and by $N = \sum_i n_i$ the total number of events. The binning is chosen such that n_i is large enough to consider that the Poisson distribution approaches a Gaussian (in practice, $n_i \geq 5$ is acceptable). If the p.d.f. of x is $f(x;\boldsymbol{\theta})$, the expected probability at the best-fit parameter $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, for an event to appear in bin i, is

$$p_i(\hat{\boldsymbol{\theta}}) = \frac{1}{C} \int_{x_i^{\text{low}}}^{x_i^{\text{up}}} f(x;\hat{\boldsymbol{\theta}}) \, \mathrm{d}x, \tag{1.65}$$

where x_i^{low} and x_i^{up} are the bin limits and C is a normalisation constant. When the histogram range covers all possible values of x, $C = 1$ since the p.d.f. is normalised; otherwise, $C = \sum_i \int_{x_i^{\text{low}}}^{x_i^{\text{up}}} f(x;\hat{\boldsymbol{\theta}}) \, \mathrm{d}x$. In practice, the integral in Eq. (1.65) is often approximated with $\Delta x_i f(x_i^c;\hat{\boldsymbol{\theta}})$, where $\Delta x_i = x_i^{\text{up}} - x_i^{\text{low}}$ and x_i^c is the value at the bin centre. When the total number of events is fixed to N (thus, we only check the shape of the distribution, not the normalisation), we can calculate the following χ^2:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left[n_i - N p_i(\hat{\boldsymbol{\theta}})\right]^2}{N p_i(\hat{\boldsymbol{\theta}})}. \tag{1.66}$$

It follows a chi-square probability function (1.60), with $(n - m - 1)$ degrees of freedom (the -1 is coming from the constraint on the total number of events). Then, the p-value (1.63) provides a criterion to evaluate the quality of the fit. Alternative χ^2 can be built to test the goodness of fit. They are described extensively in Baker and Cousins (1984).

The χ^2-tests that use the p.d.f. (1.60) are limited by the underlying assumption that the measurements have Gaussian errors. It is generally a good assumption, but not always. For instance, binned data can contain nearly empty bins, where the Poisson distribution with a low mean cannot be approximated by a Gaussian. Also, the real distribution of errors could look more or less Gaussian, but it actually has long tails that cannot be described by a Gaussian. A possible workaround is to find a function that transforms the data into another variable $\boldsymbol{x}' = f(\boldsymbol{x})$ whose errors are more Gaussian-like. The χ^2-test is then performed on $\boldsymbol{x}'$. If it is not possible, one can always generate the expected χ^2 distribution with simulated pseudo-experiments from which a p-value is calculated.

Claiming a Discovery and Setting Limits

Let us imagine that we observe a bump in the distribution of some data, which might be a hint of a new particle. To quantify whether the bump is significant, we introduce two hypotheses: the null hypothesis, H_0, corresponding to the absence of new physics (i.e., the known background), considered to be true by default; and the alternative hypothesis, H_1, where a new signal is allowed on top of the usual background. The joint p.d.f., $f(\boldsymbol{x}|H)$, of the measured variables $\boldsymbol{x} = x_1, \ldots, x_n$ depends on the hypothesis. Let us assume that a test statistic (which depends on the data), $t(\boldsymbol{x})$, that discriminates H_0 from H_1 can be found. It depends on $\boldsymbol{x}$, and so is itself a random variable described by its own p.d.f., $f_t(t|H)$. The two hypotheses H_0 and H_1 can be discriminated if $f_t(t|H_0)$ and $f_t(t|H_1)$ are significantly different. Suppose that by convention, $t(\boldsymbol{x})$ tends to have larger values when H_1 is true. For

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

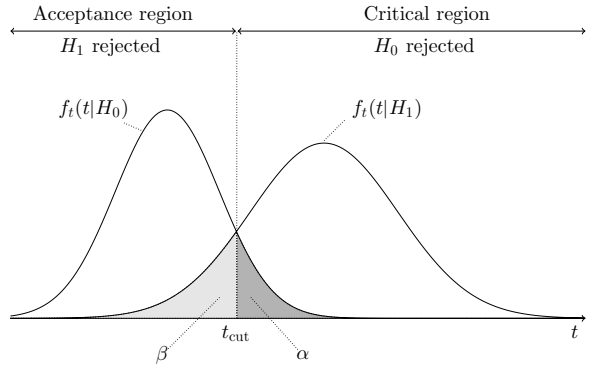


Fig. 1.16 Probability distribution of the test statistic t for the null hypothesis H_0 and the alternative H_1. The critical and acceptance regions are delimited by t_{cut}, which determines the values of the significance level, α, and the power of test $1 - \beta$ (see text).

this purpose, a value t_{cut} is chosen that delimits two regions: the so-called *critical region*, when $t > t_{\text{cut}}$, and the *acceptance region*, when $t < t_{\text{cut}}$. If the H_0 hypothesis is true, the probability to get a value t in the critical region is

$$\int_{t_{\text{cut}}}^{\infty} f_t(t|H_0)\ \mathrm{d}t = \alpha, \tag{1.67}$$

while the corresponding probability if H_1 is true is

$$\int_{t_{\text{cut}}}^{\infty} f_t(t|H_1)\ \mathrm{d}t = 1 - \int_{-\infty}^{t_{\text{cut}}} f_t(t|H_1)\ \mathrm{d}t = 1 - \beta. \tag{1.68}$$

The quantities α and β are represented in Fig 1.16 along the different regions. Testing the null hypothesis H_0 against the alternative H_1 consists in setting a significance level α (for instance, $\alpha = 1\%$), which determines the value of t_{cut}, and thus of β. The quantity $1 - \alpha$ is usually called the *confidence level* and is denoted CL, while $1 - \beta$ is called the power of the test because if $1 - \beta$ is large, the alternative hypothesis is well separated from the null hypothesis. The lower the values of α and β, or equivalently the larger the confidence level and the power, the better the test statistic is at discriminating between the two hypotheses. If the observed value of the statistic test, t_{obs}, falls in the critical region, i.e. $t_{\text{obs}} > t_{\text{cut}}$, H_0 is considered rejected. Hence, α is the probability of rejecting H_0, despite being true: it is the probability of a false discovery claim. On the other hand, β is the probability of rejecting H_1, although it is true (and thus of not rejecting H_0, despite being false): it is the probability of missing a discovery when there should be one. In the rare cases where H_0 and H_1 are the only possible hypotheses that could lead to the same data, rejecting one hypothesis is equivalent to accepting the other. In most cases, it is not equivalent because of possible alternative hypotheses not considered.

In order to claim a discovery of a new signal on top of the known background, the two hypotheses $H_0 = H_{\text{bkg}}$ and $H_1 = H_{\text{sig+bkg}}$ are then tested. The hypothesis $H_{\text{sig+bkg}}$ usually depends on a parameter of interest that we wish to measure. It is typically the cross section, σ_{sig}, of the new signal. In high energy physics, the convention is to claim

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

a discovery, or *observation*, if H_{bkg} is rejected with a significance level as low as $\alpha = 2.87 \times 10^{-7}$, equivalent to the one-sided tail probability of $Z = 5$ standard deviations of the unit Gaussian.[25] One simply says that the significance level is 5σ. As before the p-value (the observed level of significance) of the hypothesis H is

$$p\text{-value} = \int_{t_{\text{obs}}}^{\infty} f_t(t|H)\ \mathrm{d}t. \tag{1.69}$$

Hence, when $H_0 = H_{\text{bkg}}$ is rejected, it implies

$$p_{\text{bkg}} = \int_{t_{\text{obs}}}^{\infty} f_t(t|H_{\text{bkg}})\ \mathrm{d}t < \alpha. \tag{1.70}$$

If only a significance level of 3σ (i.e., $Z = 3$ and $\alpha = 1.35 \times 10^{-3}$) can be reached, it is not considered a discovery, but rather *evidence*, in the accepted jargon. Discoveries are rare, but evidence regularly appears before vanishing with the acquisition of more data. One of the reasons for this is the difficulty of correctly assessing the measurement errors, which can result from poorly modelled long tails, and thus a poor description of the tail of the probability distribution of the test statistic. So before getting too excited by an evidence, it is wiser to keep this in mind. In order to convince oneself of the discovery or the evidence, many additional elements are usually provided. For instance, one gives a confidence interval estimate of σ_{sig}, the parameter of interest under $H_{\text{sig+bkg}}$. Also, a simulation of the signal is often used to evaluate the expected p-value under the H_{bkg} hypothesis. A reasonable agreement between the expected p-value and the measured one in Eq. (1.70) strengthens the confidence in the result.

When Eq. (1.70) is not satisfied, i.e. when $p_{\text{bkg}} > \alpha$, we fail to reject H_{bkg} at the significance level, α. However, it does not imply that we have no sensitivity at all to values of σ_{sig} under $H_{\text{sig+bkg}}$. In other words, some values may be excluded and others may not. Another test is then needed to find the excluded values. It consists in testing the hypotheses $H_{\text{sig+bkg}}$ for a given value, σ_{sig}^0, against H_{bkg}. The hypothesis $H_0 = H_{\text{sig+bkg}}{:}\sigma_{\text{sig}}^0$ is excluded at the significance level, α, if its p-value satisfies

$$p_{\text{sig+bkg}} = \int_{t_{\text{obs}}}^{\infty} f_t(t|H_{\text{sig+bkg}}{:}\sigma_{\text{sig}}^0)\ \mathrm{d}t < \alpha.$$

One varies the value of σ_{sig}^0 and performs the test for each value to obtain the region of excluded σ_{sig}. For a particle-production rate or a cross section, since it is assumed that a new production mechanism can only add more events than what is expected with the background only, the excluded region is an interval of the form $[\sigma_{\text{sig}}^{\text{lim}}, +\infty]$. The standard p-value threshold for exclusion in high energy physics is $\alpha = 0.05$, i.e. 95% confidence level.

Sometimes, the number of background events may fluctuate downwards, leading to the exclusion of $H_{\text{sig+bkg}}$, although there is no sensitivity to distinguish $H_{\text{sig+bkg}}$ from H_{bkg}. In

[25] if $p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\ \mathrm{d}x$, then $Z = \Phi^{-1}(1-p)$, where $\Phi^{-1}(x)$ is the inverse function of the cumulative distribution of the unit Gaussian, i.e. $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x'^2/2}\ \mathrm{d}x'$. Hence, $\alpha = 2.87 \times 10^{-7} = \int_5^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\ \mathrm{d}x = 1 - \Phi(5)$.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

extreme cases, one may even obtain unphysical $\sigma_{\rm sig}^{\rm lim} < 0$. To prevent those cases, instead of using the p-value, $p_{\rm sig+bkg}$, the high energy physics community often uses the ratio

$$\mathrm{CL}_s = \frac{p_{\rm sig+bkg}}{1 - p_{\rm bkg}}$$

to set exclusion limits. Requiring $\mathrm{CL}_s < \alpha$ is more stringent than $p_{\rm sig+bkg} < \alpha$ since CL_s is necessarily larger than $p_{\rm sig+bkg}$. It follows that CL_s is more robust against unphysical exclusions.

One may wonder how to choose the test statistic $t(\boldsymbol{x})$. In simple analyses, it is natural to use, for instance, the number of events for counting analyses, or the distribution of an observable such as the reconstructed energy of a particle. In more complicated analyses, when the likelihood function is known, a widely used test statistic to establish discovery or exclusion limits is based on the *likelihood ratio*.[26] With $\hat{\sigma}_{\rm sig} > 0$ denoting the best fit value of the cross section of the signal hypothesis $H_{\rm sig+bkg}$ (the one maximising the likelihood), this ratio inspired from Eq. (1.54) is defined as

$$q(\sigma_{\rm sig}) = \frac{L(\sigma_{\rm sig})}{L(\hat{\sigma}_{\rm sig})}. \tag{1.71}$$

Notice that the value $\sigma_{\rm sig} = 0$ corresponds to hypothesis $H_{\rm bkg}$. For large samples, the quantity $-2 \ln q(\sigma_{\rm sig})$ is asymptotically distributed as χ^2 with, in our example of Eq. (1.71), one degree of freedom. It thus becomes easy to construct confidence intervals, as suggested by Feldman and Cousins (1998). The likelihood ratio also allows the inclusion of systematic errors as nuisance parameters that must be fitted from the data. This approach (called profile likelihood ratio) is beyond the scope of this introduction. It is detailed in Cowan et al. (2011).

Problems

1.1 **Coulomb's law.** Consider an electric charge density distribution $\rho(\boldsymbol{r})$ creating the electric field $\boldsymbol{E}$. Starting with Maxwell's equation $\boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho/\epsilon_0$, integrate the equation over a volume V, and show that

$$\oint_S \boldsymbol{E} \cdot \mathrm{d}S = Q/\epsilon_0,$$

where S is the surface encapsulating the volume V and Q is the charge contained in V. Assuming a static point-like particle with a charge q_0 located at $\boldsymbol{0}$, show that the electric field at $\boldsymbol{r}$ is

$$\boldsymbol{E} = \frac{q_0}{4\pi\epsilon_0 |\boldsymbol{r}|^2} \frac{\boldsymbol{r}}{|\boldsymbol{r}|}.$$

Conclude that the force felt by a charge q at $\boldsymbol{r}$ is given by Coulomb's law,

$$\boldsymbol{f} = \frac{q\, q_0}{4\pi\epsilon_0 |\boldsymbol{r}|^2} \frac{\boldsymbol{r}}{|\boldsymbol{r}|}.$$

[26] The popularity of the likelihood ratio is coming from the Neyman–Pearson lemma that states that the likelihood ratio between two hypotheses represents the optimal test statistic.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

1.2 **Yukawa potential.** Consider the general equation

$$(\nabla^2 - \lambda^2) f(\boldsymbol{r}) = -\delta(\boldsymbol{r}),$$

where λ is a real number. Using the Fourier transform $\tilde{f}(\boldsymbol{k}) = \iiint f(\boldsymbol{r}) e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}\, \mathrm{d}\boldsymbol{r}$ and its inverse transform $f(\boldsymbol{r}) = 1/(2\pi)^3 \iiint \tilde{f}(\boldsymbol{k}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}}\, \mathrm{d}\boldsymbol{k}$, show that the solution of the equation is

$$f(\boldsymbol{r}) = \frac{1}{(2\pi)^3} \iiint \frac{e^{i\boldsymbol{k}\cdot\boldsymbol{r}}}{|\boldsymbol{k}|^2 + \lambda^2}\, \mathrm{d}\boldsymbol{k}.$$

Using polar coordinates and orienting the z-axis of the reference frame along the $\boldsymbol{r}$ direction, check that

$$f(\boldsymbol{r}) = \frac{1}{2\pi^2} \frac{1}{|\boldsymbol{r}|} \int_0^\infty \frac{\sin(|\boldsymbol{k}||\boldsymbol{r}|)}{|\boldsymbol{k}|^2 + \lambda^2} |\boldsymbol{k}|\, d|\boldsymbol{k}|.$$

Using a contour integration in the complex plane (if you are not familiar with this technique, read Appendix F), show that

$$f(\boldsymbol{r}) = \frac{e^{-\lambda|\boldsymbol{r}|}}{4\pi|\boldsymbol{r}|}.$$

Deduce the expression of the electrostatic potential, Eq. (1.2), and the Yukawa potential, Eq. (1.1).

1.3 Let I_v be the vertical flux of muons. Given that the flux intensities coming from the zenith angle, θ, is given by $I(\theta) = I_v \cos^2\theta$, show that the muon flux from the sky collected by a horizontal detector should be $I_v\pi/2$.

1.4 Playing with relativistic formulas that are recalled in Section 2.2, check Eq. (1.9).

1.5 Without loss of generality, let $\boldsymbol{B}$ be a static magnetic field along the z-axis. Check that the power from the Lorentz force is zero and deduce that γ is a constant of the motion. Conclude that

$$\gamma m \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = q\boldsymbol{v} \times \boldsymbol{B}.$$

Show that the solution of this equation yields the following coordinates of the trajectory:

$$\begin{aligned} x(t) &= x(0) + v_y(0)/\omega + [v_x(0)\sin(\omega t) - v_y(0)\cos(\omega t)]/\omega \\ y(t) &= y(0) - v_x(0)/\omega + [v_y(0)\sin(\omega t) + v_x(0)\cos(\omega t)]/\omega \\ z(t) &= z(0) + v_z(0) \times t, \end{aligned}$$

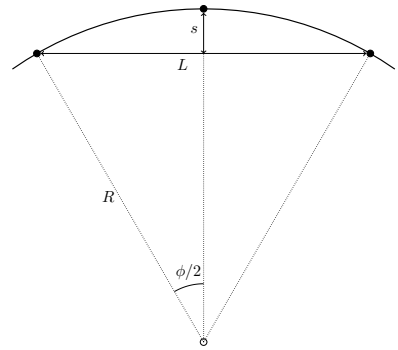
with $\omega = q|\boldsymbol{B}|/(\gamma m)$. One could introduce the variable $u = x + iy$ to ease the calculation. Finally, verify that the bending radius in the (x, y) plane satisfies Eq. (1.27).

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press

1.6 Sagitta determination.

The momentum measurement is related to the determination of the sagitta, s (image on the right). Imagine that three points are used to determine the particle track. In the approximation where the angle ϕ is small (or equivalently $s \ll L$), show that

$$s \simeq R\frac{\phi^2}{8} = \frac{L^2}{8R}.$$



Using Eq. (1.30) with $\delta = 0.3$ and $|q| = 1$, deduce that the transverse momentum is

$$|\boldsymbol{p}_\perp| = \frac{0.3|\boldsymbol{B}|L^2}{8s}.$$

Hence, conclude that the uncertainty is given by $\sigma(|\boldsymbol{p}_\perp|)/|\boldsymbol{p}_\perp| = \sigma(s)\, 8|\boldsymbol{p}_\perp|/(0.3BL^2)$.

1.7 Weighted sum and likelihood. Let r_i be the result of the ith measurement and σ_{r_i} its uncertainty. The r_i random variables, $i = 1, \ldots, n$, are assumed to follow a Gaussian p.d.f. $G(r_i; \mu, \sigma_{r_i})$. Apply the maximum likelihood method to evaluate the parameter μ. Show that the minimisation of the negative log-likelihood yields the estimator

$$\hat{\mu} = \frac{1}{\sum_i w_i} \sum_{i=1}^{N} w_i r_i,$$

with $w_i = 1/\sigma_{r_i}^2$.

1.8 Particle lifetime and least squares. The least-squares method can be solved analytically when the data are described by a linear function in the parameters to estimate. We will adapt it to the determination of the lifetime τ of the nuclei constituting a radioactive source. A detector records its activity during consecutive intervals of $\Delta t = 20$ ns:

Interval, k	1	2	3	4	5
Number of counts, n_i	2 797	1 241	570	264	128

The number of atoms of the source at a time t is given by $N(t) = N_0 \exp(-\gamma t)$, with $\gamma = 1/\tau$ ($t = 0$ is the beginning of data acquisition).

1. Show that the number of decays (i.e., counts) between t and $t + \Delta t$ satisfies $\ln[N_{\text{decay}}(t)] = -\gamma t + \alpha$, with $\alpha = \ln[N_0(1 - \exp(-\gamma \Delta t))]$.
2. Assuming that the number of counts n_k is Gaussian distributed, with a standard deviation $\sigma_k = \sqrt{n_k}$ (justified by their large values), what standard deviation should be used for $\ln(n_k)$?
3. Apply the least-squares method to the variable $\ln(n_k)$ and show that the estimator of γ is

$$\hat{\gamma} = \frac{1}{\Delta t} \frac{\left(\sum_k n_k \ln n_k\right)\left(\sum_k n_k(k-1)\right) - \left(\sum_k n_k\right)\left(\sum_k n_k \ln n_k(k-1)\right)}{\left(\sum_k n_k\right)\left(\sum_k n_k(k-1)^2\right) - \left(\sum_k n_k(k-1)\right)^2},$$

where k varies from 1 to 5. Deduce the value of τ and compare it to the true value $\tau = 25$ ns.

https://doi.org/10.1017/9781009171595.002 Published online by Cambridge University Press