2012 European School of High-Energy Physics

La Pommeraye, Anjou, France
6 – 19 June 2012

Proceedings
Editors: C. Grojean
M. Mulders
Abstract

The European School of High-Energy Physics is intended to give young physicists an introduction to the theoretical aspects of recent advances in elementary particle physics. These proceedings contain lecture notes on the Standard Model of electroweak interactions, quantum chromodynamics, flavour physics, physics beyond the Standard Model, neutrino physics, and cosmology.
Preface

The twentieth event in the series of the European School of High-Energy Physics took place in La Pommeraye, Anjou, France, from 6 to 19 June 2012. It was organized jointly by CERN, Geneva, Switzerland, and JINR, Dubna, Russia, with support from CNRS/IN2P3, CEA/Irfu, the Angers-Loire local administration, and the commune of La Pommeraye. The local organization team was chaired by Sophie Henrot-Versille. The other members of the local committee were: Valérie Brouillard, Caroline Collard, Fabrice Couderc, Philippe Crochet, Jean-Baptiste de Vivie de Régie, Emi Kou, François Le Diberder, Julie Malclès and Benoit Viaud.

A total of 98 students coming from 34 different countries attended the school, mainly from member states of CERN and/or JINR, but also a few from other regions. The participants were generally students in experimental High-Energy Physics in the final years of work towards their PhDs.

The School was hosted at the Jardins de l’Anjou complex and conference centre in La Pommeraye, about 30 km from the city of Angers and 330 km South-West from Paris. According to the tradition of the school, the students shared twin rooms mixing participants of different nationalities.

A total of 33 lectures were complemented by daily discussion sessions led by six discussion leaders. The students displayed their own research work in the form of posters in an evening session in the first week, and the posters stayed on display until the end of the School. Each discussion group carried out a collaborative project, studying in detail the analysis from a published paper from one of the LHC experiments; a summary was presented by a student representative of each group in an evening session in the second week of the School. The full scientific programme was arranged in the on-site conference facilities.

Our thanks go to the local-organization team and, in particular, to Sophie Henrot-Versille for all her work and assistance in preparing the School, on both scientific and practical matters, and for her presence during the event. Our thanks also go to the efficient and friendly hotel management and staff who assisted the School organizers and the participants in many ways.

Very great thanks are due to the lecturers and discussion leaders for their active participation in the School and for making the scientific programme so stimulating. The students, who in turn manifested their good spirits during two intense weeks, undoubtedly appreciated listening to and discussing with the teaching staff of world renown.

We would like to express our appreciation to Professor Rolf Heuer, Director General of CERN, and Professor Victor Matveev, Director General of JINR, for their lectures on the scientific programmes of the two organizations and for discussing with the School participants.

In addition to the rich scientific programme, the participants enjoyed numerous sports and leisure activities in and around the Jardins de l’Anjou complex. Particularly noteworthy were the very nice excursions to the city of Angers, to the Atlantic coast at La Baule-Escoublac and the city of Nantes, and the cruise on the river Loire with a visit to the Chateau des Vaults where the participants were able to taste fine wines from the local region.

Various Outreach activities took place around the School, including contacts with the local lycée as well as a public lecture in Angers. We were very grateful to the mayor and town council of La Pommeraye for their interest and support. We would also like to thank Jacques Martino, Director of IN2P3, and Laurent Serin, Deputy Director of IN2P3, for visiting the School.

We are very grateful to Mrs Hélène Haller and Mrs Tatyana Donskova for their untiring efforts in the lengthy preparations for and the day-to-day operation of the School. Their continuous care of the participants and their needs during the School was highly appreciated. The success of the School was to a large extent due to the students themselves. Their poster session was very well prepared and highly appreciated, their group projects were a great success, and throughout the School they participated actively during the lectures, in the discussion sessions and in the different activities and excursions. Finally, one should not forget the show that they prepared and presented following the farewell banquet.

Nick Ellis
(On behalf of the Organizing Committee)
### People in the photograph

1. Jeffrey WETTER
2. Philippe CROCHET
3. Anton KRACHENKO
4. Stephen COLE
5. Adam Edward BARTON
6. Kristian GREGersen
7. Peter Lundba ROSENDAHL
8. Lukas Fritz MARTI
9. Brett JACKSON
10. Stefan WAYAND
11. Claudio HELLER
12. Homero MARTINEZ
13. Stepan OBRAZTsov
14. Simon AKAR
15. Rosa SIMONIELLO
16. Christoph WASICKI
17. Bastian BEISCHER
18. Jonas WEICHERT
19. Christophe WEICHERT
20. Aleksey KHUDYAKOV
21. Maurizio PIERINI
22. Garoe GONZALEZ PARRA
23. Dmitry TSIRKOV
24. Jesper Roy CHRISTIANSEN
25. Joe PRICE
26. Yossof ESHAQ
27. Theo BRISTOW
28. Christian BARTH
29. Anthony HAWKINGS
30. Xiaoxiao WANG
31. Driss CHARFEDDINE
32. Olivier BONDU
33. Gaetano BARONE
34. Federico BERTOLUCCI
35. Martin RYBAR
36. Mathieu PERRIN-TERRIN
37. Jean-Baptiste DE VIVIE
38. Lukas BAENI
39. Jean ILOPOULOS
40. Philipp ELLER
41. Nick ELLIS
42. Giacomo SNIDERO
43. Cristiano ALPIGIANI
44. Fabrice COUDERC
45. Felix BACHMAIR
46. Konstantin KANISHCHEV
47. Basil SCHNEIDER
48. Leonid SERKIN
49. Daniel NOONAN
50. Joana MACHADO MIGUENS
51. Robert EBER

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>52.</td>
<td>Tamara VAZQUEZ SCHROEDER</td>
</tr>
<tr>
<td>53.</td>
<td>Caitlin MALONE</td>
</tr>
<tr>
<td>54.</td>
<td>Bugra BILIN</td>
</tr>
<tr>
<td>55.</td>
<td>Robert CLARKE</td>
</tr>
<tr>
<td>56.</td>
<td>Daniele ZANZI</td>
</tr>
<tr>
<td>57.</td>
<td>Sarah BERANEK</td>
</tr>
<tr>
<td>58.</td>
<td>Maria EROFEEEVA</td>
</tr>
<tr>
<td>59.</td>
<td>Laura ZAMBELLI</td>
</tr>
<tr>
<td>60.</td>
<td>Francesco RUBBO</td>
</tr>
<tr>
<td>61.</td>
<td>Francesco COSTANZA</td>
</tr>
<tr>
<td>62.</td>
<td>Simone Federico BRAZZALE</td>
</tr>
<tr>
<td>63.</td>
<td>Ivan PRADO LONGHI</td>
</tr>
<tr>
<td>64.</td>
<td>Dean LAMBERT</td>
</tr>
<tr>
<td>65.</td>
<td>Andrej ABRUZOV</td>
</tr>
<tr>
<td>66.</td>
<td>Itamar ROTH</td>
</tr>
<tr>
<td>67.</td>
<td>Artur SHAIKHIEV</td>
</tr>
<tr>
<td>68.</td>
<td>Alexey GLADYSHEV</td>
</tr>
<tr>
<td>69.</td>
<td>Alexander GRAMOLIN</td>
</tr>
<tr>
<td>70.</td>
<td>Helene HALLER</td>
</tr>
<tr>
<td>71.</td>
<td>Yury STEPANENKO</td>
</tr>
<tr>
<td>72.</td>
<td>Sofia Maria CONSONNI</td>
</tr>
<tr>
<td>73.</td>
<td>Oleksii IVANYTSKYI</td>
</tr>
<tr>
<td>74.</td>
<td>Joany MANIARRÉS RAMOS</td>
</tr>
<tr>
<td>75.</td>
<td>Effychia-Sofia PROTOPADAKI</td>
</tr>
<tr>
<td>76.</td>
<td>Lene BRYGENMARK</td>
</tr>
<tr>
<td>77.</td>
<td>Meng XIAO</td>
</tr>
<tr>
<td>78.</td>
<td>Nickolas MC COLL</td>
</tr>
<tr>
<td>79.</td>
<td>Harrison PROSPER</td>
</tr>
<tr>
<td>80.</td>
<td>Nadja STROBBE</td>
</tr>
<tr>
<td>81.</td>
<td>Olga KOCHEBINA</td>
</tr>
<tr>
<td>82.</td>
<td>Valentin KNUNZ</td>
</tr>
<tr>
<td>83.</td>
<td>Boris BULANEK</td>
</tr>
<tr>
<td>84.</td>
<td>Miroslav HAVRANEK</td>
</tr>
<tr>
<td>85.</td>
<td>Karol ADAMCZYK</td>
</tr>
<tr>
<td>86.</td>
<td>Jared STURDY</td>
</tr>
<tr>
<td>87.</td>
<td>Nathaniel ODELL</td>
</tr>
<tr>
<td>88.</td>
<td>James PEARCE</td>
</tr>
<tr>
<td>89.</td>
<td>Andrés Guillermo DELANNOY</td>
</tr>
<tr>
<td>90.</td>
<td>Hendrik JANSSEN</td>
</tr>
<tr>
<td>91.</td>
<td>Adrian CHITAN</td>
</tr>
<tr>
<td>92.</td>
<td>Vieri ADRIAN</td>
</tr>
<tr>
<td>93.</td>
<td>Valerio BORTOLOTTO</td>
</tr>
<tr>
<td>94.</td>
<td>Faig AHMADOV</td>
</tr>
<tr>
<td>95.</td>
<td>Eva RIBEZL</td>
</tr>
<tr>
<td>96.</td>
<td>Valerie LANG</td>
</tr>
<tr>
<td>97.</td>
<td>Samuel Ross MEEHAN</td>
</tr>
<tr>
<td>98.</td>
<td>Houry KEOSHKERIAN</td>
</tr>
<tr>
<td>99.</td>
<td>Petr GALLUS</td>
</tr>
<tr>
<td>100.</td>
<td>Itzebelt SANTOYO CASTILLO</td>
</tr>
<tr>
<td>101.</td>
<td>Tatyana DONSKOVA</td>
</tr>
<tr>
<td>102.</td>
<td>Martijn MULDERS</td>
</tr>
</tbody>
</table>
PHOTOGRAPHS (MONTAGE)
Contents

Preface
  N. Ellis ................................................................. vii
Photograph of participants ................................................. viii
Photographs (montage) ......................................................... xi
Introduction to the Standard Model of the Electro-Weak Interactions
  J. Iliopoulos ............................................................ 1
A brief Introduction to Modern Amplitude Methods
  L. J. Dixon ............................................................... 31
Flavor Physics and CP Violation
  G. Isidori ................................................................. 69
Is (Low-Energy) SUSY still Alive?
  A.V. Gladyshev and D.I. Kazakov .................................... 107
Neutrino Physics
  G. Barenboim ........................................................... 161
LHC Results–Highlights
  G. Rolandi ............................................................... 181
Practical Statistics for Particle Physicists
  H. B. Prosper .......................................................... 195
Cosmology and Gravitation: the grand Scheme for High-Energy Physics
  P. Binétruy ............................................................. 217
Organizing Committee ......................................................... 293
Local Organizing Committee ................................................. 293
List of Lecturers .............................................................. 293
List of Discussion Leaders .................................................. 293
List of Students .............................................................. 294
List of Posters ............................................................... 295
Introduction to the Standard Model of Electro-Weak Interactions

J. Iliopoulos
Laboratoire de Physique Théorique de L’Ecole Normale Supérieure, Paris, France

Abstract
These lectures notes cover the basic ideas of gauge symmetries and the phenomenon of spontaneous symmetry breaking which are used in the construction of the Standard Model of the Electro-Weak Interactions.

1 Introduction
These are the notes of a set of four lectures which I gave at the 2012 CERN Summer School of Particle Physics. They were supposed to serve as introductory material to more specialised lectures. The students were mainly young graduate students in Experimental High Energy Physics. They were supposed to be familiar with the phenomenology of Particle Physics and to have a working knowledge of Quantum Field Theory and the techniques of Feynman diagrams. The lectures were concentrated on the physical ideas underlying the concept of gauge invariance, the mechanism of spontaneous symmetry breaking and the construction of the Standard Model. Although the methods of computing higher order corrections and the theory of renormalisation were not discussed at all in the lectures, the general concept of renormalisable versus non-renormalisable theories was supposed to be known.

The plan of the notes follows that of the lectures with five sections:

- a brief summary of the phenomenology of the electromagnetic and the weak interactions;
- gauge theories, Abelian and non-Abelian;
- spontaneous symmetry breaking;
- the step-by-step construction of the Standard Model;
- the Standard Model and experiment.

It is only in the last part that the notes differ from the actual lectures, because I took into account the recent evidence for a Higgs boson.

It is generally admitted that progress in physics occurs when an unexpected experimental result contradicts the established theoretical beliefs. As Feynman has put it “progress in Physics is to prove yourself wrong as soon as possible.” This has been the rule in the past, but there are exceptions. The construction of the Standard Model is one of them. In the late sixties weak interactions were well described by the Fermi current x current theory and there was no compelling experimental reason to want to change it. Its problems were theoretical. It was only a phenomenological model which, in the technical language, was non-renormalisable. In practice this means that any attempt to compute higher order corrections in the standard perturbation theory would give meaningless divergent results. So, the motivation was aesthetic rather than experimental, it was the search of mathematical consistency and theoretical elegance. In fact, at the beginning, the data did not seem to support the theoretical speculations. Although the history of these ideas is a fascinating subject, I decided not to follow the historical evolution. I start instead from the experimental data known at present and show that they point unmistakably to what is known as the Standard Model. It is only at the last section where I recall its many experimental successes.

2 Phenomenology of the electro-weak interactions: a reminder

2.1 The elementary particles
The notion of an “Elementary Particle” is not well-defined in High Energy Physics. It evolves with time following the progress in the experimental techniques which, by constantly increasing the resolution...
Table 1: This table shows our present ideas on the structure of matter. Quarks and gluons do not exist as free particles and the graviton has not yet been observed.

<table>
<thead>
<tr>
<th>Quanta of radiation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong Interactions</td>
<td>Eight gluons</td>
</tr>
<tr>
<td>Electromagnetic Interactions</td>
<td>Photon ((\gamma))</td>
</tr>
<tr>
<td>Weak Interactions</td>
<td>Bosons (W^+), (W^-), (Z^0)</td>
</tr>
<tr>
<td>Gravitational Interactions</td>
<td>Graviton (?)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Matter particles</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Leptons</td>
<td>Quarks</td>
</tr>
<tr>
<td>1st Family</td>
<td>(\nu_e, e^-)</td>
</tr>
<tr>
<td>2nd Family</td>
<td>(\nu_\mu, \mu^-)</td>
</tr>
<tr>
<td>3rd Family</td>
<td>(\nu_\tau, \tau^-)</td>
</tr>
<tr>
<td></td>
<td>(u_a, d_a, a = 1, 2, 3)</td>
</tr>
<tr>
<td></td>
<td>(c_a, s_a, a = 1, 2, 3)</td>
</tr>
<tr>
<td></td>
<td>(t_a, b_a, a = 1, 2, 3)</td>
</tr>
</tbody>
</table>

Higgs boson

power of our observations, have shown that systems which were believed to be “elementary”, are in fact composite out of smaller constituents. So, in the last century we went through the chain: molecules \(\rightarrow\) atoms \(\rightarrow\) electrons + nuclei \(\rightarrow\) electrons + protons + neutrons \(\rightarrow\) electrons + quarks \(\rightarrow\) ???

There is no reason to believe that there is an end in this series and, even less, that this end has already been reached. Table 1 summarises our present knowledge.

Some remarks concerning this table:

- all interactions are produced by the exchange of virtual quanta. For the strong, electromagnetic and weak interactions they are vector (spin-one) fields, while the graviton is assumed to be a tensor, spin-two field. We shall see in these lectures that this property is well understood in the framework of gauge theories;

- the constituents of matter appear to be all spin one-half particles. They are divided into quarks, which are hadrons, and “leptons” which have no strong interactions. No deep explanation is known neither for their number, (why three families?), nor for their properties, such as their quantum numbers. We shall come back to this point when we discuss the gauge theory models. In the framework of some theories going beyond the Standard Model, such as supersymmetric theories, among the matter constituents we can find particles of zero spin;

- each quark species, called “flavour”, appears under three forms, often called “colours” (no relation with the ordinary sense of the words);

- quarks and gluons do not appear as free particles. They form a large number of bound states, the hadrons. This property of “confinement” is one of the deep unsolved problems in Particle Physics;

- quarks and leptons seem to fall into three distinct groups, or “families”. No deep explanation is known;

- the mathematical consistency of the theory, known as “the cancellation of the triangle anomalies”, requires that the sum of all electric charges inside any family is equal to zero. This property has a strong predictive power.

2.2 The electromagnetic interactions

All experimental data are well described by a simple interaction Lagrangian in which the photon field interacts with a current built out of the fields of charged particles.
For the spinor matter fields of the Table the current takes the simple form:

\[ j^\mu(x) = \sum_i q_i \bar{\Psi}_i(x) \gamma^\mu \Psi_i(x) \]  

where \( q_i \) is the charge of the field \( \Psi_i \) in units of \( e \).

This simple Lagrangian has some remarkable properties, all of which are verified by experiment:

- \( j \) is a vector current. The interaction conserves separately \( P, \ C\) and \( T \);
- the current is diagonal in flavour space;
- more complex terms, such as \( j^\mu(x) j_\mu(x), \partial A(x) \bar{\Psi}(x) \Psi(x), \ldots \) are absent, although they do not seem to be forbidden by any known property of the theory. All these terms, as well as all others we can write, share one common property: In a four-dimensional space-time, their canonical dimension is larger than four. We can easily show that the resulting quantum field theory is non-renormalisable. For some reason, Nature does not like non-renormalisable theories.

Quantum electrodynamics, the quantum field theory described by the Lagrangian (1) and supplemented with the programme of renormalisation, is one of the most successful physical theories. Its agreement with experiment is spectacular. For years it was the prototype for all other theories. The Standard Model is the result of the efforts to extend the ideas and methods of the electromagnetic interactions to all other forces in physics.

2.3 The weak interactions

They are mediated by massive vector bosons. When the Standard Model was proposed their very existence, as well as their number, was unknown. But today we know that there exist three, two which are electrically charged and one neutral: \( W^+, W^- \) and \( Z^0 \). Like the photon, their couplings to matter are described by current operators:

\[ L_i \sim V_\mu(x) j^\mu(x) ; \ V_\mu : W^+, W^-; Z^0_\mu \] 

where the weak currents are again bi-linear in the fermion fields: \( \bar{\Psi} \ldots \Psi \). Depending on the corresponding vector boson, we distinguish two types of weak currents: the charged current, coupled to \( W^+ \) and \( W^- \) and the neutral current coupled to \( Z^0 \). They have different properties:

- **The charged current:**
  - contains only left-handed fermion fields:
    \[ j_\mu \sim \bar{\Psi}_L \gamma_\mu \Psi_L \sim \bar{\Psi}_L (1 + \gamma_5) \Psi \]  
  - it is non-diagonal in the quark flavour space;
  - the coupling constants are complex.

- **The neutral current:**
  - contains both left- and right-handed fermion fields:
    \[ j_\mu \sim C_L \bar{\Psi}_L \gamma_\mu \Psi_L + C_R \bar{\Psi}_R \gamma_\mu \Psi_R \]  
  - it is diagonal in the quark flavour space.

With these currents weak interactions have some properties which differ from those of the electromagnetic ones:
• weak interactions violate $P$, $C$ and $T$;
• contrary to the photon, the weak vector bosons are self-coupled. The nature of these couplings is predicted theoretically in the framework of gauge theories and it has been determined experimentally;
• a new element has been added recently to the experimental landscape: We have good evidence for the existence of a new particle, compatible with what theorists have called the Higgs boson, although its properties have not yet been studied in detail.

It is this kind of interactions that the Standard Model is supposed to describe.

3 Gauge symmetries

3.1 The concept of symmetry

In Physics the concept of a Symmetry follows from the assumption that a certain quantity is not measurable. As a result the equations of motion should not depend on this quantity. We know from the general properties of Classical Mechanics that this implies the existence of conserved quantities. This relation between symmetries and conservation laws, epitomised by Noether’s theorem, has been one of the most powerful tools in deciphering the properties of physical theories.

Some simple examples are given by the symmetries of space and time. The assumption that the position of the origin of the coordinate system is not physically measurable implies the invariance of the equations under space translations and the conservation of momentum. In the same way we obtain the conservation laws of energy (time translations) and angular momentum (rotations). We can also distinguish between symmetries under continuous transformations, such as translations and rotations, and discrete ones, such as space, or time, inversions. Noether’s theorem applies to the first.

All these symmetries of space and time are geometrical in the common sense of the word, easy to understand and visualise. During the last century we were led to consider two abstractions, each one of which has had a profound influence in our way of thinking the fundamental interactions. Reversing the chronological order, we shall introduce first the idea of internal symmetries and, second, that of local or gauge symmetries.

3.2 Internal symmetries

We call internal symmetries those whose transformation parameters do not affect the point of space and time $x$. The concept of such symmetries can be presented already in classical physics, but it becomes natural in quantum mechanics and quantum field theory. The simplest example is the phase of the wave function. We know that it is not a measurable quantity, so the theory must be invariant under a change of phase. This is true for both relativistic or non-relativistic quantum mechanics. The equations of motion (Dirac or Schrödinger), as well as the normalisation condition, are invariant under the transformation:

$$\Psi(x) \rightarrow e^{i\theta} \Psi(x)$$  \hspace{1cm} (6)

The transformation leaves the space-time point invariant, so it is an internal symmetry. Through Noether’s theorem, invariance under (6) implies the conservation of the probability current.

The phase transformation (6) corresponds to the Abelian group $U(1)$. In 1932 Werner Heisenberg enlarged the concept to a non-Abelian symmetry with the introduction of isospin. The assumption is that strong interactions are invariant under a group of $SU(2)$ transformations in which the proton and the neutron form a doublet $N(x)$:

$$N(x) = \left( \begin{array}{c} p(x) \\ n(x) \end{array} \right) \hspace{1cm} ; \hspace{1cm} N(x) \rightarrow e^{i\vec{\tau} \cdot \vec{\theta}} N(x)$$  \hspace{1cm} (7)
where \( \vec{\tau} \) are proportional to the Pauli matrices and \( \vec{\theta} \) are the three angles of a general rotation in a three dimensional Euclidean space. Again, the transformations do not apply on the points of ordinary space.

Heisenberg’s iso-space is three dimensional, isomorphic to our physical space. With the discovery of new internal symmetries the idea was generalised to multi-dimensional internal spaces. The space of Physics, i.e. the space in which all symmetry transformations apply, became an abstract mathematical concept with non-trivial geometrical and topological properties. Only a part of it, the three-dimensional Euclidean space, is directly accessible to our senses.

### 3.3 Gauge symmetries

The concept of a local, or gauge, symmetry was introduced by Albert Einstein in his quest for the theory of General Relativity\(^1\). Let us come back to the example of space translations, as shown in Fig. 1.

The figure shows that, if \( A \) is the trajectory of a free particle in the \((x,y,z)\) system, its image, \( A' \), is also a possible trajectory of a free particle in the new system. The dynamics of free particles is invariant under space translations by a constant vector. It is a *global* invariance, in the sense that the parameter \( \vec{a} \) is independent of the space-time point \( x \). Is it possible to extend this invariance to a *local* one, namely one in which \( \vec{a} \) is replaced by an arbitrary function of \( x \); \( \vec{a}(x) \)? One calls usually the transformations in which the parameters are functions of the space-time point \( x \) *gauge transformations*\(^2\). There may be various, essentially aesthetic, reasons for which one may wish to extend a global invariance to a gauge one. In physical terms, one may argue that the formalism should allow for a local definition of the origin of the coordinate system, since the latter is an unobservable quantity. From the mathematical point of view local transformations produce a much richer and more interesting structure. Whichever one’s motivations may be, physical or mathematical, it is clear that the free particle dynamics is not invariant under translations in which \( \vec{a} \) is replaced by \( \vec{a}(x) \). This is shown schematically in Fig. 2.

We see that no free particle, in its right minds, would follow the trajectory \( A'' \). This means that, for \( A'' \) to be a trajectory, the particle must be subject to external forces. Can we determine these forces? The question sounds purely geometrical without any obvious physical meaning, so we expect a mathematical answer with no interest for Physics. The great surprise is that the resulting theory which is invariant under local translations turns out to be Classical General Relativity, one of the four fundamental forces in Nature. Gravitational interactions have such a geometric origin. In fact, the mathematical formulation of Einstein’s original motivation to extend the Principle of Equivalence to accelerated frames, is precisely the requirement of local invariance. Historically, many mathematical techniques which are used in today’s gauge theories were developed in the framework of General Relativity.

---

\(^1\)It is also present in classical electrodynamics if one considers the invariance under the change of the vector potential \( A_\mu(x) \rightarrow A_\mu(x) - \partial_\mu \theta(x) \) with \( \theta \) an arbitrary function, but before the introduction of quantum mechanics, this aspect of the symmetry was not emphasised.

\(^2\)This strange terminology is due to Hermann Weyl. In 1918 he attempted to enlarge diffeomorphisms to local scale transformations and he called them, correctly, *gauge transformations*. The attempt was unsuccessful, but, when in 1929 he developed the theory for the Dirac electron, although the theory is no more scale invariant, he still used the term gauge invariance, a term which has survived ever since.
The gravitational forces are not the only ones which have a geometrical origin. Let us come back to the example of the quantum mechanical phase. It is clear that neither the Dirac nor the Schrödinger equation are invariant under a local change of phase $\theta(x)$. To be precise, let us consider the free Dirac Lagrangian:

$$
\mathcal{L} = \bar{\Psi}(x)(i\partial - m)\Psi(x)
$$

(8)

It is not invariant under the transformation:

$$
\Psi(x) \rightarrow e^{i\theta(x)}\Psi(x)
$$

(9)

The reason is the presence of the derivative term in (8) which gives rise to a term proportional to $\partial_\mu \theta(x)$. In order to restore invariance, one must modify (8), in which case it will no longer describe a free Dirac field; invariance under gauge transformations leads to the introduction of interactions. Both physicists and mathematicians know the answer to the particular case of (8): one introduces a new field $A_\mu(x)$ and replaces the derivative operator $\partial_\mu$ by a “covariant derivative” $D_\mu$ given by:

$$
D_\mu = \partial_\mu + ieA_\mu
$$

(10)

where $e$ is an arbitrary real constant. $D_\mu$ is called “covariant” because it satisfies

$$
D_\mu [e^{i\theta(x)}\Psi(x)] = e^{i\theta(x)}D_\mu\Psi(x)
$$

(11)

valid if, at the same time, $A_\mu(x)$ undergoes the transformation:

$$
A_\mu(x) \rightarrow A_\mu(x) - \frac{1}{e}\partial_\mu\theta(x)
$$

(12)

The Dirac Lagrangian density becomes now:

$$
\mathcal{L} = \bar{\Psi}(x)(iD - m)\Psi(x) = \bar{\Psi}(x)(i\partial - eA - m)\Psi(x)
$$

(13)

It is invariant under the gauge transformations (9) and (12) and describes the interaction of a charged spinor field with an external electromagnetic field! Replacing the derivative operator by the covariant derivative turns the Dirac equation into the same equation in the presence of an external electromagnetic field. Electromagnetic interactions admit the same geometrical interpretation\(^3\). We can complete the picture by including the degrees of freedom of the electromagnetic field itself and add to (13) the corresponding Lagrangian density. Again, gauge invariance determines its form uniquely and we are led to the well-known result:

\(^3\)The same applies to the Schrödinger equation. In fact, this was done first by V. Fock in 1926, immediately after Schrödinger’s original publication.
\[ \mathcal{L} = -\frac{1}{4} F_{\mu\nu}(x) F^{\mu\nu}(x) + \bar{\Psi}(x) (iD / -m) \Psi(x) \]  
(14)

with

\[ F_{\mu\nu}(x) = \partial_\mu A_\nu(x) - \partial_\nu A_\mu(x) \]  
(15)

The constant \( e \) we introduced is the electric charge, the coupling strength of the field \( \Psi \) with the electromagnetic field. Notice that a second field \( \Psi' \) will be coupled with its own charge \( e' \).

Let us summarise: We started with a theory invariant under a group \( U(1) \) of global phase transformations. The extension to a local invariance can be interpreted as a \( U(1) \) symmetry at each point \( x \). In a qualitative way we can say that gauge invariance induces an invariance under \( U(1)^\infty \). We saw that this extension, a purely geometrical requirement, implies the introduction of new interactions. The surprising result here is that these “geometrical” interactions describe the well-known electromagnetic forces.

The extension of the formalism of gauge theories to non-Abelian groups is not trivial and was first discovered by trial and error. Here we shall restrict ourselves to internal symmetries which are simpler to analyse and they are the ones we shall apply to particle physics outside gravitation.

Let us consider a classical field theory given by a Lagrangian density \( \mathcal{L} \). It depends on a set of \( N \) fields \( \psi^i(x), \; i = 1, \ldots, r \) and their first derivatives. The Lorentz transformation properties of these fields will play no role in this discussion. We assume that the \( \psi^i \)'s transform linearly according to an \( r \)-dimensional representation, not necessarily irreducible, of a compact, simple, Lie group \( G \) which does not act on the space-time point \( x \).

\[ \Psi = \begin{pmatrix} \psi^1 \\ \vdots \\ \psi^r \end{pmatrix} \hspace{1cm} \Psi(x) \to U(\omega) \Psi(x) \hspace{1cm} \omega \in G \]  
(16)

where \( U(\omega) \) is the matrix of the representation of \( G \). In fact, in these lectures we shall be dealing only with perturbation theory and it will be sufficient to look at transformations close to the identity in \( G \).

\[ \Psi(x) \to e^{i\Theta} \Psi(x) \hspace{1cm} \Theta = \sum_{a=1}^{m} \theta^a T^a \]  
(17)

where the \( \theta^a \)'s are a set of \( m \) constant parameters, and the \( T^a \)'s are \( m \times m \) matrices representing the \( m \) generators of the Lie algebra of \( G \). They satisfy the commutation rules:

\[ [T^a, T^b] = i f^{abc} T^c \]  
(18)

The \( f \)'s are the structure constants of \( G \) and a summation over repeated indices is understood. The normalisation of the structure constants is usually fixed by requiring that, in the fundamental representation, the corresponding matrices of the generators \( t^a \) are normalised such as:

\[ \text{Tr} \left( t^a t^b \right) = \frac{1}{2} \delta^{ab} \]  
(19)

The Lagrangian density \( \mathcal{L}(\Psi, \partial \Psi) \) is assumed to be invariant under the global transformations (17) or (16). As was done for the Abelian case, we wish to find a new \( \mathcal{L} \), invariant under the corresponding gauge transformations in which the \( \theta^a \)'s of (17) are arbitrary functions of \( x \). In the same qualitative sense, we look for a theory invariant under \( G^\infty \). This problem, stated the way we present it here, was first solved by trial and error for the case of \( SU(2) \) by C.N. Yang and R.L. Mills in 1954. They gave the underlying physical motivation and these theories are called since “Yang-Mills theories”. The steps are
direct generalisations of the ones followed in the Abelian case. We need a gauge field, the analogue of the electromagnetic field, to transport the information contained in (17) from point to point. Since we can perform \( m \) independent transformations, the number of generators in the Lie algebra of \( G \), we need \( m \) gauge fields \( A^a_\mu(x) \), \( a = 1, \ldots, m \). It is easy to show that they belong to the adjoint representation of \( G \). Using the matrix representation of the generators we can cast \( A^a_\mu(x) \) into an \( r \times r \) matrix:

\[
A^a_\mu(x) = \sum_{a=1}^{m} A^a_\mu(x) T^a
\]

(20)

The covariant derivatives can now be constructed as:

\[
D_\mu = \partial_\mu + ig A_\mu
\]

(21)

with \( g \) an arbitrary real constant. They satisfy:

\[
D_\mu e^{i\theta(x)} \Psi(x) = e^{i\theta(x)} D_\mu \Psi(x)
\]

(22)

provided the gauge fields transform as:

\[
A^a_\mu(x) \rightarrow e^{i\theta(x)} A^a_\mu(x) e^{-i\theta(x)} + \frac{i}{g} \left( \partial_\mu e^{i\theta(x)} \right) e^{-i\theta(x)}
\]

(23)

The Lagrangian density \( \mathcal{L}(\Psi, D\Psi) \) is invariant under the gauge transformations (17) and (23) with an \( x \)-dependent \( \Theta \), if \( \mathcal{L}(\Psi, \partial_\mu \Psi) \) is invariant under the corresponding global ones (16) or (17). As was done with the electromagnetic field, we can include the degrees of freedom of the new gauge fields by adding to the Lagrangian density a gauge invariant kinetic term. It turns out that it is slightly more complicated than \( F_{\mu\nu} \) of the Abelian case. Yang and Mills computed it for \( SU(2) \) but, in fact, it is uniquely determined by geometry plus some obvious requirements, such as absence of higher order derivatives. The result is given by:

\[
G_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - ig [A_\mu, A_\nu]
\]

(24)

The full gauge-invariant Lagrangian can now be written as:

\[
\mathcal{L}_{\text{inv}} = -\frac{1}{2} \text{Tr} G_{\mu\nu} G^{\mu\nu} + \mathcal{L}(\Psi, D\Psi)
\]

(25)

By convention, in (24) the matrix \( A \) is taken to be:

\[
A_\mu = A^a_\mu T^a
\]

(26)

where we recall that the \( T^a \)'s are the matrices representing the generators in the fundamental representation. It is only with this convention that the kinetic term in (25) is correctly normalised. In terms of the component fields \( A^a_\mu \), \( G_{\mu\nu} \) reads:

\[
G_{\mu\nu} = G^a_{\mu\nu} T^a
\]

(27)

\[
G^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g f^{abc} A^b_\mu A^c_\nu
\]

Under a gauge transformation \( G_{\mu\nu} \) transforms like a member of the adjoint representation.

\[
G_{\mu\nu}(x) \rightarrow e^{i\theta(x)} T^a \quad G^a_{\mu\nu}(x) \rightarrow e^{-i\theta(x)} T^a
\]

(28)

This completes the construction of the gauge invariant Lagrangian. We add some remarks:

8
as it was the case with the electromagnetic field, the Lagrangian (25) does not contain terms proportional to $A_\mu A^\mu$. This means that, under the usual quantisation rules, the gauge fields describe massless particles;

- since $G_{\mu\nu}$ is not linear in the fields $A_\mu$, the $G^2$ term in (25), besides the usual kinetic term which is bilinear in the fields, contains tri-linear and quadri-linear terms. In perturbation theory they will be treated as coupling terms whose strength is given by the coupling constant $g$. In other words, the non-Abelian gauge fields are self-coupled while the Abelian (photon) field is not. A Yang-Mills theory, containing only gauge fields, is still a dynamically rich quantum field theory while a theory with the electromagnetic field alone is a trivial free theory;

- the same coupling constant $g$ appears in the covariant derivative of the fields $\Psi$ in (21). This simple consequence of gauge invariance has an important physical application: if we add another field $\Psi'$, its coupling strength with the gauge fields will still be given by the same constant $g$. Contrary to the Abelian case studied before, if electromagnetism is part of a non-Abelian simple group, gauge invariance implies charge quantisation;

- the above analysis can be extended in a straightforward way to the case where the group $G$ is the product of simple groups $G = G_1 \times ... \times G_n$. The only difference is that one should introduce $n$ coupling constants $g_1, ..., g_n$, one for each simple factor. Charge quantisation is still true inside each subgroup, but charges belonging to different factors are no more related;

- the situation changes if one considers non semi-simple groups, where one, or more, of the factors $G_i$ is Abelian. In this case the associated coupling constants can be chosen different for each field and the corresponding Abelian charges are not quantised.

As we alluded to above, gauge theories have a deep geometrical meaning. In order to get a better understanding of this property without entering into complicated issues of differential geometry, it is instructive to consider a reformulation of the theory replacing the continuum of space-time with a four dimensional Euclidean lattice. We can do that very easily. Let us consider, for simplicity, a lattice with hypercubic symmetry. The space-time point $x_\mu$ is replaced by:

$$x_\mu \rightarrow n_\mu a$$

where $a$ is a constant length, (the lattice spacing), and $n_\mu$ is a $d$-dimensional vector with components $n_\mu = (n_1, n_2, ..., n_d)$ which take integer values $0 \leq n_\mu \leq N_\mu$. $N_\mu$ is the number of points of our lattice in the direction $\mu$. The total number of points, i.e. the volume of the system, is given by $V \sim \prod_{\mu=1}^d N_\mu$. The presence of $a$ introduces an ultraviolet, or short distance, cut-off because all momenta are bounded from above by $2\pi/a$. The presence of $N_\mu$ introduces an infrared, or large distance cut-off because the momenta are also bounded from below by $2\pi/Na$, where $N$ is the maximum of $N_\mu$. The infinite volume continuum space is recovered at the double limit $a \rightarrow 0$ and $N_\mu \rightarrow \infty$.

The dictionary between quantities defined in the continuum and the corresponding ones on the lattice is easy to establish (we take the lattice spacing $a$ equal to one):

- a field $\Psi(x) \Rightarrow \Psi_n$

where the field $\Psi$ is an $r$-component column vector as in equation (16);

- a local term such as $\bar{\Psi}(x)\Psi(x) \Rightarrow \bar{\Psi}_n\Psi_n$;

- a derivative $\partial_\mu \Psi(x) \Rightarrow (\Psi_n - \Psi_{n+\mu})$;

where $n + \mu$ should be understood as a unit vector joining the point $n$ with its nearest neighbour in the direction $\mu$;

- the kinetic energy term$^4$ $\bar{\Psi}(x)\partial_\mu \Psi(x) \Rightarrow \bar{\Psi}_n\Psi_n - \bar{\Psi}_n\Psi_{n+\mu}$

$^4$We write here the expression for spinor fields which contain only first order derivatives in the kinetic energy. The extension to scalar fields with second order derivatives is obvious.
We may be tempted to write similar expressions for the gauge fields, but we must be careful with the way gauge transformations act on the lattice. Let us repeat the steps we followed in the continuum: Under gauge transformations a field transforms as:

- gauge transformations
  \[ \Psi(x) \rightarrow e^{i\Theta(x)}\Psi(x) \Rightarrow \Psi_n \rightarrow e^{i\Theta_n}\Psi_n \]

All local terms of the form \( \overline{\Psi}_n \Psi_n \) remain invariant but the part of the kinetic energy which couples fields at neighbouring points does not.

- the kinetic energy
  \[ \overline{\Psi}_n \Psi_n + \mu \rightarrow \overline{\Psi}_n e^{-i\Theta_n} e^{i\Theta_{n+\mu}} \Psi_n + \mu \]

which shows that we recover the problem we had with the derivative operator in the continuum. In order to restore invariance we must introduce a new field, which is an \( r \)-by-\( r \) matrix, and which has indices \( n \) and \( n + \mu \). We denote it by \( U_{n,n+\mu} \) and we shall impose on it the constraint \( U_{n,n+\mu} = U_{n+\mu,n}^{-1} \). Under a gauge transformations, \( U \) transforms as:

\[ U_{n,n+\mu} \rightarrow e^{i\Theta_n} U_{n,n+\mu} e^{-i\Theta_{n+\mu}} \quad (30) \]

With the help of this gauge field we write the kinetic energy term with the covariant derivative on the lattice as:

\[ \overline{\Psi}_n U_{n,n+\mu} \Psi_{n+\mu} \quad (31) \]

which is invariant under gauge transformations.

\( U \) is an element of the gauge group but we can show that, at the continuum limit and for an infinitesimal transformation, it reproduces correctly \( A_\mu \), which belongs to the Lie algebra of the group. Notice that, contrary to the field \( \Psi \), \( U \) does not live on a single lattice point, but it has two indices, \( n \) and \( n + \mu \), in other words it lives on the oriented link joining the two neighbouring points. We see here that the mathematicians are right when they do not call the gauge field “a field” but “a connection”.

In order to finish the story we want to obtain an expression for the kinetic energy of the gauge field, the analogue of \( \text{Tr} G_{\mu\nu}(x) G^{\mu\nu}(x) \), on the lattice. As for the continuum, the guiding principle is gauge invariance. Let us consider two points on the lattice \( n \) and \( m \). We shall call a path \( p_{n,m} \) on the lattice a sequence of oriented links which join continuously the two points. Consider next the product of the gauge fields \( U \) along all the links of the path \( p_{n,m} \):

\[ P^{(p)}(n, m) = \prod_p U_{n,n+\mu}...U_{m-\nu,m} \quad (32) \]

Using the transformation rule (30), we see that \( P^{(p)}(n, m) \) transforms as:

\[ P^{(p)}(n, m) \rightarrow e^{i\Theta_n} P^{(p)}(n, m) e^{-i\Theta_m} \quad (33) \]

It follows that if we consider a closed path \( c = p_{n,n} \) the quantity \( \text{Tr} P^{(c)} \) is gauge invariant. The simplest closed path for a hypercubic lattice has four links and it is called plaquette. The correct form of the Yang-Mills action on the lattice can be written in terms of the sum of \( \text{Tr} P^{(c)} \) over all plaquettes.

4 Spontaneous symmetry breaking

Since gauge theories appear to predict the existence of massless gauge bosons, when they were first proposed they did not seem to have any direct application to particle physics outside electromagnetism. It is this handicap which plagued gauge theories for many years. In this section we shall present a seemingly unrelated phenomenon which, however, will turn out to provide the answer.
An infinite system may exhibit the phenomenon of phase transitions. It often implies a reduction in the symmetry of the ground state. A field theory is a system with an infinite number of degrees of freedom, so, it is not surprising that field theories may also show the phenomenon of phase transitions. Indeed, in many cases, we encounter at least two phases:

- **the unbroken, or, the Wigner phase**: The symmetry is manifest in the spectrum of the theory whose excitations form irreducible representations of the symmetry group. For a gauge theory the vector gauge bosons are massless and belong to the adjoint representation. But we have good reasons to believe that, for non-Abelian gauge theories, a strange phenomenon occurs in this phase: all physical states are singlets of the group. All non-singlet states, such as those corresponding to the gauge fields, are supposed to be confined, in the sense that they do not appear as physically realisable asymptotic states;

- **the spontaneously broken phase**: Part of the symmetry is hidden from the spectrum. For a gauge theory, some of the gauge bosons become massive and appear as physical states.

It is this kind of phase transition that we want to study in this section.

### 4.1 An example from classical mechanics

A very simple example is provided by the problem of the bent rod. Let a cylindrical rod be charged as in Fig. 3. The problem is obviously symmetric under rotations around the z-axis. Let z measure the distance from the basis of the rod, and X(z) and Y(z) give the deviations, along the x and y directions respectively, of the axis of the rod at the point z from the symmetric position. For small deflections the equations of elasticity take the form:

\[ IE \frac{d^4 X}{dz^4} + F \frac{d^2 X}{dz^2} = 0 \quad \text{;} \quad IE \frac{d^4 Y}{dz^4} + F \frac{d^2 Y}{dz^2} = 0 \]  

where \( I = \pi R^4 / 4 \) is the moment of inertia of the rod and \( E \) is the Young modulus. It is obvious that the system (34) always possesses a symmetric solution \( X = Y = 0 \). However, we can also look for asymmetric solutions of the general form: \( X = A + Bz + C \sin k z + D \cos k z \) with \( k^2 = F / EI \), which satisfy the boundary conditions \( X = X'' = 0 \) at \( z = 0 \) and \( z = l \). We find that such solutions exist, \( X = C \sin k z \), provided \( kl = n \pi \); \( n = 1, \ldots \). The first such solution appears when \( F \) reaches a critical value \( F_{cr} \) given by:

\[ F_{cr} = \frac{\pi^2 EI}{l^2} \]  

The appearance of these solutions is already an indication of instability and, indeed, a careful
study of the stability problem proves that the non-symmetric solutions correspond to lower energy. From that point Eqs. (34) are no longer valid, because they only apply to small deflections, and we must use the general equations of elasticity. The result is that this instability of the symmetric solution occurs for all values of $F$ larger than $F_{cr}$.

What has happened to the original symmetry of the equations? It is still hidden in the sense that we cannot predict in which direction in the $x - y$ plane the rod is going to bend. They all correspond to solutions with precisely the same energy. In other words, if we apply a symmetry transformation (in this case a rotation around the $z$-axis) to an asymmetric solution, we obtain another asymmetric solution which is degenerate with the first one.

We call such a symmetry “spontaneously broken”, and in this simple example we see all its characteristics:

- there exists a critical point, i.e. a critical value of some external quantity which we can vary freely, (in this case the external force $F$; in several physical systems it is the temperature) which determines whether spontaneous symmetry breaking will take place or not. Beyond this critical point:
  - the symmetric solution becomes unstable;
  - the ground state becomes degenerate.

There exist a great variety of physical systems, both in classical and quantum physics, exhibiting spontaneous symmetry breaking, but we will not describe any other one here. The Heisenberg ferromagnet is a good example to keep in mind, because we shall often use it as a guide, but no essentially new phenomenon appears outside the ones we saw already. Therefore, we shall go directly to some field theory models.

### 4.2 A simple field theory model

Let $\phi(x)$ be a complex scalar field whose dynamics is described by the Lagrangian density:

$$\mathcal{L}_1 = (\partial_\mu \phi)(\partial^\mu \phi^*) - M^2 \phi \phi^* - \lambda (\phi \phi^*)^2$$

(36)

where $\mathcal{L}_1$ is a classical Lagrangian density and $\phi(x)$ is a classical field. No quantisation is considered for the moment. The Lagrangian (36) is invariant under the group $U(1)$ of global transformations:

$$\phi(x) \to e^{i\theta} \phi(x)$$

(37)

To this invariance corresponds the current $j_\mu \sim \phi \partial_\mu \phi^* - \phi^* \partial_\mu \phi$ whose conservation can be verified using the equations of motion.

We are interested in the classical field configuration which minimises the energy of the system. We thus compute the Hamiltonian density given by

$$\mathcal{H}_1 = (\partial_0 \phi)(\partial_t \phi^*) + (\partial_t \phi)(\partial_0 \phi^*) + V(\phi)$$

(38)

$$V(\phi) = M^2 \phi \phi^* + \lambda (\phi \phi^*)^2$$

(39)

The first two terms of $\mathcal{H}_1$ are positive definite. They can only vanish for $\phi = \text{constant}$. Therefore, the ground state of the system corresponds to $\phi = \text{constant} = \text{minimum of } V(\phi)$. $V$ has a minimum only if $\lambda > 0$. In this case the position of the minimum depends on the sign of $M^2$. (Notice that we are still studying a classical field theory and $M^2$ is just a parameter. One should not be misled by the notation into thinking that $M$ is a “mass” and $M^2$ is necessarily positive).
For $M^2 > 0$ the minimum is at $\phi = 0$ (symmetric solution, shown in the left side of Fig. 4), but for $M^2 < 0$ there is a whole circle of minima at the complex $\phi$-plane with radius $v = (-M^2/2\lambda)^{1/2}$ (Fig. 4, right side). Any point on the circle corresponds to a spontaneous breaking of (37).

We see that:

- the critical point is $M^2 = 0$;
- for $M^2 > 0$ the symmetric solution is stable;
- for $M^2 < 0$ spontaneous symmetry breaking occurs.

Let us choose $M^2 < 0$. In order to reach the stable solution we translate the field $\phi$. It is clear that there is no loss of generality by choosing a particular point on the circle, since they are all obtained from any given one by applying the transformations (37). Let us, for convenience, choose the point on the real axis in the $\phi$-plane. We thus write:

$$\phi(x) = \frac{1}{\sqrt{2}} [v + \psi(x) + i\chi(x)]$$  \hspace{1cm} (40)

Bringing (40) in (36) we find

$$\mathcal{L}_1(\phi) \rightarrow \mathcal{L}_2(\psi, \chi) = \frac{1}{2} (\partial_\mu \psi)^2 + \frac{1}{2} (\partial_\mu \chi)^2 - \frac{1}{2} (2\lambda v^2) \psi^2 - \lambda v \psi (\psi^2 + \chi^2) - \frac{\lambda}{4} (\psi^2 + \chi^2)^2$$  \hspace{1cm} (41)

Notice that $\mathcal{L}_2$ does not contain any term proportional to $\chi^2$, which is expected since $V$ is locally flat in the $\chi$ direction. A second remark concerns the arbitrary parameters of the theory. $\mathcal{L}_1$ contains two such parameters, a mass $M$ and a dimensionless coupling constant $\lambda$. In $\mathcal{L}_2$ we have again the coupling constant $\lambda$ and a new mass parameter $v$ which is a function of $M$ and $\lambda$. It is important to notice that, although $\mathcal{L}_2$ contains also trilinear terms, their coupling strength is not a new parameter but is proportional to $v\lambda$. $\mathcal{L}_2$ is still invariant under the transformations with infinitesimal parameter $\theta$: 

Fig. 4: The potential $V(\phi)$ with $M^2 \geq 0$ (left) and $M^2 < 0$ (right)
\[ \delta \psi = -\theta \chi \quad ; \quad \delta \chi = \theta \psi + \theta v \]  

(42)

to which corresponds a conserved current

\[ j_\mu \sim \psi \partial_\mu \chi - \chi \partial_\mu \psi + v \partial_\mu \chi \]  

(43)

The last term, which is linear in the derivative of \( \chi \), is characteristic of the phenomenon of spontaneous symmetry breaking.

It should be emphasised here that \( L_1 \) and \( L_2 \) are completely equivalent Lagrangians. They both describe the dynamics of the same physical system and a change of variables, such as (40), cannot change the physics. However, this equivalence is only true if we can solve the problem exactly. In this case we shall find the same solution using either of them. However, we do not have exact solutions and we intend to apply perturbation theory, which is an approximation scheme. Then the equivalence is no longer guaranteed and, in fact, perturbation theory has much better chances to give sensible results using one language rather than the other. In particular, if we use \( L_1 \) as a quantum field theory and we decide to apply perturbation theory taking, as the unperturbed part, the quadratic terms of \( L_1 \), we immediately see that we shall get nonsense. The spectrum of the unperturbed Hamiltonian would consist of particles with negative square mass, and no perturbation corrections, at any finite order, could change that. This is essentially due to the fact that, in doing so, we are trying to calculate the quantum fluctuations around an unstable solution and perturbation theory is just not designed to do so. On the contrary, we see that the quadratic part of \( L_2 \) gives a reasonable spectrum; thus we hope that perturbation theory will also give reasonable results. Therefore we conclude that our physical system, considered now as a quantum system, consists of two interacting scalar particles, one with mass \( m_\psi^2 = \frac{2\lambda}{\nu^2} \) and the other with \( m_\chi = 0 \). We believe that this is the spectrum we would have found also starting from \( L_1 \), if we could solve the dynamics exactly.

The appearance of a zero-mass particle in the quantum version of the model is an example of a general theorem due to J. Goldstone: To every generator of a spontaneously broken symmetry there corresponds a massless particle, called the Goldstone particle. This theorem is just the translation, into quantum field theory language, of the statement about the degeneracy of the ground state. The ground state of a system described by a quantum field theory is the vacuum state, and you need massless excitations in the spectrum of states in order to allow for the degeneracy of the vacuum.

4.3 Gauge symmetries

In this section we want to study the consequences of spontaneous symmetry breaking in the presence of a gauge symmetry. We shall find a very surprising result. When combined together the two problems, namely the massless gauge bosons on the one hand and the massless Goldstone bosons on the other, will solve each other. It is this miracle that we want to present here. We start with the Abelian case.

We look at the model of the previous section in which the \( U(1) \) symmetry (37) has been promoted to a local symmetry with \( \theta \rightarrow \theta(x) \). As we explained already, this implies the introduction of a massless vector field, which we can call the “photon” and the interactions are obtained by replacing the derivative operator \( \partial_\mu \) by the covariant derivative \( D_\mu \) and adding the photon kinetic energy term:

\[ L_1 = -\frac{1}{4} F_{\mu \nu}^2 + |(\partial_\mu + i e A_\mu)\phi|^2 - M^2 \phi \phi^* - \lambda (\phi \phi^*)^2 \]  

(44)

\( L_1 \) is invariant under the gauge transformation:

\[ \phi(x) \rightarrow e^{i\theta(x)} \phi(x) \quad ; \quad A_\mu \rightarrow A_\mu - \frac{1}{e} \partial_\mu \theta(x) \]  

(45)
The same analysis as before shows that for $\lambda > 0$ and $M^2 < 0$ there is a spontaneous breaking of the $U(1)$ symmetry. Replacing (40) into (44) we obtain:

$$L_1 \rightarrow L_2 = -\frac{1}{4} F_{\mu\nu}^2 + \frac{e^2 v^2}{2} A_\mu^2 + e v A_\mu \partial^\mu \chi + \frac{1}{2} (\partial_\mu \psi)^2 + \frac{1}{2} (\partial_\mu \chi)^2 - \frac{1}{2} (2\lambda v^2) \psi^2 + \ldots$$

(46)

where the dots stand for coupling terms which are at least trilinear in the fields.

The surprising term is the second one which is proportional to $A_\mu^2$. It looks as though the photon has become massive. Notice that (46) is still gauge invariant since it is equivalent to (44). The gauge transformation is now obtained by replacing (40) into (45):

$$\psi(x) \rightarrow \cos \theta(x) \psi(x) + v \sin \theta(x) \chi(x) - v$$

$$\chi(x) \rightarrow \cos \theta(x) \chi(x) + \sin \theta(x) \psi(x) + v$$

(47)

$$A_\mu \rightarrow A_\mu - \frac{1}{e} \partial_\mu \theta(x)$$

This means that our previous conclusion, that gauge invariance forbids the presence of an $A_\mu^2$ term, was simply wrong. Such a term can be present, only the gauge transformation is slightly more complicated; it must be accompanied by a translation of the field.

The Lagrangian (46), if taken as a quantum field theory, seems to describe the interaction of a massive vector particle ($A_\mu$) and two scalars, one massive ($\psi$) and one massless ($\chi$). However, we can see immediately that something is wrong with this counting. A warning is already contained in the non-diagonal term between $A_\mu$ and $\partial^\mu \chi$. Indeed, the perturbative particle spectrum can be read from the Lagrangian only after we have diagonalised the quadratic part. A more direct way to see the trouble is to count the apparent degrees of freedom before and after the translation:

- Lagrangian (44):
  (i) one massless vector field: 2 degrees
  (ii) one complex scalar field: 2 degrees
  Total: 4 degrees

- Lagrangian (46):
  (i) one massive vector field: 3 degrees
  (ii) two real scalar fields: 2 degrees
  Total: 5 degrees

Since physical degrees of freedom cannot be created by a simple change of variables, we conclude that the Lagrangian (46) must contain fields which do not create physical particles. This is indeed the case, and we can exhibit a transformation which makes the unphysical fields disappear. Instead of parametrising the complex field $\phi$ by its real and imaginary parts, let us choose its modulus and its phase.

---

5The terminology here is misleading. As we pointed out earlier, any field theory, considered as a dynamical system, is a system with an infinite number of degrees of freedom. For example, the quantum theory of a free neutral scalar field is described by an infinite number of harmonic oscillators, one for every value of the three-dimensional momentum. Here we use the same term “degrees of freedom” to denote the independent one-particle states. We know that for a massive spin-$s$ particle we have $2s + 1$ one-particle states and for a massless particle with spin different from zero we have only two. In fact, it would have been more appropriate to talk about a $(2s + 1)$-infinity and $2$-infinity degrees of freedom, respectively.
The choice is dictated by the fact that it is a change of phase that describes the motion along the circle of the minima of the potential $V(\phi)$. We thus write:

$$\phi(x) = \frac{1}{\sqrt{2}} [v + \rho(x)] e^{i\zeta(x)/v}; \quad A_\mu(x) = B_\mu(x) - \frac{1}{ev} \partial_\mu \zeta(x)$$ (48)

In this notation, the gauge transformation (45) or (47) is simply a translation of the field $\zeta$: $\zeta(x) \rightarrow \zeta(x) + v\theta(x)$. Replacing (48) into (44) we obtain:

$$\mathcal{L}_1 \rightarrow \mathcal{L}_3 = -\frac{1}{4} B_{\mu\nu}^2 + \frac{e^2v^2}{2} B_\mu^2 + \frac{1}{2} (\partial_\mu \rho)^2 - \frac{1}{2} (2\lambda v^2) \rho^2$$

$$- \frac{\lambda}{4} \rho^4 + \frac{1}{2} e^2 B_\mu^2 (2v \rho + \rho^2)$$

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$$ (49)

The field $\zeta(x)$ has disappeared. Formula (49) describes two massive particles, a vector ($B_\mu$) and a scalar ($\rho$). It exhibits no gauge invariance, since the original symmetry $\zeta(x) \rightarrow \zeta(x) + v\theta(x)$ is now trivial.

We see that we obtained three different Lagrangians describing the same physical system. $\mathcal{L}_1$ is invariant under the usual gauge transformation, but it contains a negative square mass and, therefore, it is unsuitable for quantisation. $\mathcal{L}_2$ is still gauge invariant, but the transformation law (47) is more complicated. It can be quantised in a space containing unphysical degrees of freedom. This, by itself, is not a great obstacle and it occurs frequently. For example, ordinary quantum electrodynamics is usually quantised in a space containing unphysical degrees of freedom. This, by itself, is unsuitable for quantisation. $\mathcal{L}_3$ is no longer invariant under any kind of gauge transformation, but it exhibits clearly the particle spectrum of the theory. It contains only physical particles and they are all massive. This is the miracle that was announced earlier. Although we start from a gauge theory, the final spectrum contains massive particles only. Actually, $\mathcal{L}_3$ can be obtained from $\mathcal{L}_2$ by an appropriate choice of gauge.

The conclusion so far can be stated as follows:

In a spontaneously broken gauge theory the gauge vector bosons acquire a mass and the would-be massless Goldstone bosons decouple and disappear. Their degrees of freedom are used in order to make possible the transition from massless to massive vector bosons.

The extension to the non-Abelian case is straightforward. Let us consider a gauge group $G$ with $m$ generators and, thus, $m$ massless gauge bosons. The claim is that we can break part of the symmetry spontaneously, leaving a subgroup $H$ with $h$ generators unbroken. The $h$ gauge bosons associated to $H$ remain massless while the $m-h$ others acquire a mass. In order to achieve this result we need $m-h$ scalar degrees of freedom with the same quantum numbers as the broken generators. They will disappear from the physical spectrum and will re-appear as zero helicity states of the massive vector bosons. As previously, we shall see that one needs at least one more scalar state which remains physical.

In the remaining of this section we show explicitly these results for a general gauge group. The reader who is not interested in technical details may skip this part.

We introduce a multiplet of scalar fields $\phi_i$ which transform according to some representation, not necessarily irreducible, of $G$ of dimension $n$. According to the rules we explained in the last section, the Lagrangian of the system is given by:

$$\mathcal{L} = -\frac{1}{4} \text{Tr}(G_{\mu\nu}G^{\mu\nu}) + (D_\mu \Phi)^\dagger D^\mu \Phi - V(\Phi)$$ (50)
In component notation, the covariant derivative is, as usual, $D_\mu \phi_i = \partial_\mu \phi_i - ig^{(a)}_i T^a_\mu \phi_j$ where we have allowed for the possibility of having arbitrary coupling constants $g^{(a)}_i$ for the various generators of $G$ because we do not assume that $G$ is simple or semi-simple. $V(\Phi)$ is a polynomial in $\Phi$ invariant under $G$ of degree equal to four. As before, we assume that we can choose the parameters in $V$ such that the minimum is not at $\Phi = 0$ but rather at $\Phi = v$ where $v$ is a constant vector in the representation space of $\Phi$. $v$ is not unique. The $m$ generators of $G$ can be separated into two classes: $h$ generators which annihilate $v$ and form the Lie algebra of the unbroken subgroup $H$; and $m - h$ generators, represented in the representation of $\Phi$ by matrices $T^a$, such that $T^a v \neq 0$ and all vectors $T^a v$ are independent and can be chosen orthogonal. Any vector in the orbit of $v$, i.e. of the form $e^{iwT^a}v$ is an equivalent minimum of the potential. As before, we should translate the scalar fields $\Phi$ into orthogonal. Any vector in the orbit of $v$ and orthogonal to it, the analogue of the $\chi$ and $\psi$ fields of the previous section. We can write:

$$\Phi = i \sum_{a=1}^{m-h} \frac{\chi^a T^a v}{|T^a v|} + \sum_{b=1}^{n-m+h} \psi^b u^b + v$$

(51)

where the vectors $u^b$ form an orthonormal basis in the space orthogonal to all $T^a v$'s. The corresponding generators span the coset space $G/H$. As before, we shall show that the fields $\chi^a$ will be absorbed by the Higgs mechanism and the fields $\psi^b$ will remain physical. Note that the set of vectors $u^b$ contains at least one element since, for all $a$, we have:

$$v \cdot T^a v = 0$$

(52)

because the generators in a real unitary representation are anti-symmetric. This shows that the dimension $n$ of the representation of $\Phi$ must be larger than $m - h$ and, therefore, there will remain at least one physical scalar field which, in the quantum theory, will give a physical scalar particle $^8$.

Let us now bring in the Lagrangian (50) the expression of $\Phi$ from (51). We obtain:

$$\mathcal{L} = \frac{1}{2} \sum_{a=1}^{m-h} (\partial_\mu \chi^a)^2 + \frac{1}{2} \sum_{b=1}^{n-m+h} (\partial_\mu \psi^b)^2 - \frac{1}{4} \text{Tr}(F_{\mu \nu}F^{\mu \nu})$$

$$+ \frac{1}{2} \sum_{a=1}^{m-h} g^{(a)2}|T^a v|^2 A^a_\mu A^{\mu a} - \sum_{a=1}^{m-h} g^{(a)2} T^a v \partial_\mu \chi^a A^a_\mu - V(\Phi) + \ldots$$

(53)

where the dots stand for coupling terms between the scalars and the gauge fields. In writing (53) we took into account that $T^b v = 0$ for $b > m - h$ and that the vectors $T^a v$ are orthogonal.

The analysis that gave us Goldstone’s theorem shows that

$$\frac{\partial^2 V}{\partial \phi_k \partial \phi_l} |_{\phi = v}(T^a v) = 0$$

(54)

which shows that the $\chi$-fields would correspond to the Goldstone modes. As a result, the only mass terms which appear in $V$ in equation (53) are of the form $\psi^k M_{kl} \psi^l$ and do not involve the $\chi$-fields.

As far as the bilinear terms in the fields are concerned, the Lagrangian (53) is the sum of terms of the form found in the Abelian case. All gauge bosons which do not correspond to $H$ generators acquire a

---

^8Obviously, the argument assumes the existence of scalar fields which induce the phenomenon of spontaneous symmetry breaking. We can construct models in which the role of the latter is played by some kind of fermion-antifermion bound states and they come under the name of models with a dynamical symmetry breaking. In such models the existence of a physical spin-zero state, the analogue of the $\sigma$-particle of the chiral symmetry breaking of QCD, is a dynamical question, in general hard to answer.
mass equal to \( m_a = g^{(a)} |T^a v| \) and, through their mixing with the would-be Goldstone fields \( \chi \), develop a zero helicity state. All other gauge bosons remain massless. The \( \psi \)'s represent the remaining physical Higgs fields.

5 Building the Standard Model: a five step programme

In this section we shall construct the Standard Model of electro-weak interactions as a spontaneously broken gauge theory. We shall follow the hints given by experiment following a five step programme:

- step 1: choose a gauge group \( G \);
- step 2: choose the fields of the “elementary” particles and assign them to representations of \( G \). Include scalar fields to allow for the Higgs mechanism;
- step 3: write the most general renormalisable Lagrangian invariant under \( G \). At this stage gauge invariance is still exact and all gauge vector bosons are massless;
- step 4: choose the parameters of the Higgs potential so that spontaneous symmetry breaking occurs;
- step 5: translate the scalars and rewrite the Lagrangian in terms of the translated fields. Choose a suitable gauge and quantise the theory.

A remark: Gauge theories provide only the general framework, not a detailed model. The latter will depend on the particular choices made in steps 1) and 2).

5.1 The lepton world

We start with the leptons and, in order to simplify the presentation, we shall assume that neutrinos are massless. We follow the five steps:

- Step 1: looking at the Table of elementary particles we see that, for the combined electromagnetic and weak interactions, we have four gauge bosons, namely \( W^\pm, Z^0 \) and the photon. As we explained earlier, each one of them corresponds to a generator of the group \( G \). The only non-trivial group with four generators is \( U(2) \approx SU(2) \times U(1) \).

Following the notation which was inspired by the hadronic physics, we call \( T_i, i = 1, 2, 3 \) the three generators of \( SU(2) \) and \( Y \) that of \( U(1) \). Then, the electric charge operator \( Q \) will be a linear combination of \( T_3 \) and \( Y \). By convention, we write:

\[
Q = T_3 + \frac{1}{2} Y
\]  

The coefficient in front of \( Y \) is arbitrary and only fixes the normalisation of the \( U(1) \) generator relatively to those of \( SU(2) \)\(^7\). This ends our discussion of the first step.

- Step 2: the number and the interaction properties of the gauge bosons are fixed by the gauge group. This is no more the case with the fermion fields. In principle, we can choose any number and assign them to any representation. It follows that the choice here will be dictated by the phenomenology.

Leptons have always been considered as elementary particles. We have six leptons, however, as we noticed already, a striking feature of the data is the phenomenon of family repetition. We do not understand why Nature chooses to repeat itself three times, but the simplest way to incorporate this observation to the model is to use three times the same representations, one for each family. This leaves \( SU(2) \) doublets and/or singlets as the only possible choices. A further experimental input we shall use is the fact that the charged \( W \)'s couple only to the left-handed components of the lepton fields, contrary

\(^7\)The normalisation of the generators for non-Abelian groups is fixed by their commutation relations. That of the Abelian generator is arbitrary. The relation (55) is one choice which has only a historical value. It is not the most natural one from the group theory point of view, as you will see in the discussion concerning Grand-Unified theories.
to the photon which couples with equal strength to both right and left. These considerations lead us to
assign the left-handed components of the lepton fields to doublets of $SU(2)$.

$$\Psi_L(x) = \frac{1}{2} (1 + \gamma_5) \left( \begin{array}{c} \nu_i(x) \\ \ell_i^-(x) \end{array} \right) ; \ i = 1, 2, 3$$ (56)

where we have used the same symbol for the particle and the associated Dirac field.

The right-handed components are assigned to singlets of $SU(2)$:

$$\nu_{iR}(x) = \frac{1}{2} (1 - \gamma_5) \nu_i(x) \quad ; \quad \ell_{iR}^-(x) = \frac{1}{2} (1 - \gamma_5) \ell_i^-(x)$$ (57)

The question mark next to the right-handed neutrinos means that the presence of these fields is not
confirmed by the data. We shall drop them in this lecture, but we may come back to this point later. We
shall also simplify the notation and put $\ell_{iR}^-(x) = R_i(x)$. The resulting transformation properties under
local $SU(2)$ transformations are:

$$\Psi_L^i(x) \to e^{i\vec{\tau}\vec{\theta}(x)} \Psi_L^i(x) \quad ; \quad R_i(x) \to R_i(x)$$ (58)

with $\vec{\tau}$ the three Pauli matrices. This assignment and the $Y$ normalisation given by Eq. (55), fix also the
$U(1)$ charge and, therefore, the transformation properties of the lepton fields. For all $i$ we find:

$$Y(\Psi_L^i) = -1 \quad ; \quad Y(R_i) = -2$$ (59)

If a right-handed neutrino exists, it has $Y(\nu_{iR}) = 0$, which shows that it is not coupled to any
gauge boson.

We are left with the choice of the Higgs scalar fields and we shall choose the solution with the
minimal number of fields. We must give masses to three vector gauge bosons and keep the fourth one
massless. The latter will be identified with the photon. We recall that, for every vector boson acquiring
mass, a scalar with the same quantum numbers decouples. At the end we shall remain with at least one
physical, neutral, scalar field. It follows that the minimal number to start with is four, two charged and
two neutral. We choose to put them, under $SU(2)$, into a complex doublet:

$$\Phi = \left( \begin{array}{c} \phi^+ \\ \phi^0 \end{array} \right) \quad ; \quad \Phi(x) \to e^{i\vec{\tau}\vec{\theta}(x)} \Phi(x)$$ (60)

with the conjugate fields $\phi^-$ and $\phi^{0*}$ forming $\Phi^\dagger$. The $U(1)$ charge of $\Phi$ is $Y(\Phi) = 1$.

This ends our choices for the second step. At this point the model is complete. All further steps
are purely technical and uniquely defined.

- Step 3: What follows is straightforward algebra. We write the most general, renormalisable,
Lagrangian, involving the fields (56), (57) and (60) invariant under gauge transformations of $SU(2) \times
U(1)$. We shall also assume the separate conservation of the three lepton numbers, leaving the discussion
on the neutrino mixing to a specialised lecture. The requirement of renormalisability implies that all
terms in the Lagrangian are monomials in the fields and their derivatives and their canonical dimension
is smaller or equal to four. The result is:

$$\mathcal{L} = -\frac{1}{4} \bar{W}_{\mu\nu} \cdot \tilde{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} + |D_\mu \Phi|^2 - V(\Phi)$$

$$+ \sum_{i=1}^3 \left[ \bar{\Psi}_L^i i\partial_\mu \Psi_L^i + \bar{R}_i i\partial_\mu R_i - G_i (\bar{\Psi}_L^i R_i \Phi + h.c.) \right]$$ (61)
If we call $\vec{W}$ and $B$ the gauge fields associated to $SU(2)$ and $U(1)$ respectively, the corresponding field strengths $\vec{W}_{\mu \nu}$ and $B_{\mu \nu}$ appearing in (61) are given by (24) and (15).

Similarly, the covariant derivatives in (61) are determined by the assumed transformation properties of the fields, as shown in (21):

$$D_\mu \Psi^I_L = \left( \partial_\mu - ig \frac{\tau^I}{2} \cdot \vec{W}_\mu + ig' \frac{B_\mu}{2} \right) \Psi^I_L ; \quad D_\mu R_i = (\partial_\mu + ig' B_\mu) R_i$$

$$D_\mu \Phi = \left( \partial_\mu - ig \frac{\tau^I}{2} \cdot \vec{W}_\mu - ig' B_\mu \right) \Phi$$

(62)

The two coupling constants $g$ and $g'$ correspond to the groups $SU(2)$ and $U(1)$ respectively. The most general Higgs potential $V(\Phi)$ compatible with the transformation properties of the field $\Phi$ is:

$$V(\Phi) = \mu^2 |\Phi|^2 + \lambda (\Phi^{\dagger} \Phi)^2$$

(63)

The last term in (61) is a Yukawa coupling term between the scalar $\Phi$ and the fermions. In the absence of right-handed neutrinos, this is the most general term which is invariant under $SU(2) \times U(1)$. As usual, $h.c.$ stands for “hermitian conjugate”. $G_i$ are three arbitrary coupling constants. If right-handed neutrinos exist there is a second Yukawa term with $R_i$ replaced by $\nu_{iR}$ and $\Phi$ by the corresponding doublet proportional to $\tau^I \Phi^*, \nu^*$, where $* \mu$ means “complex conjugation”. We see that the Standard model can perfectly well accommodate a right-handed neutrino, but it couples only to the Higgs field.

A final remark: As expected, the gauge bosons $\vec{W}_\mu$ and $B_\mu$ appear to be massless. The same is true for all fermions. This is not surprising because the assumed different transformation properties of the right and left handed components forbid the appearance of a Dirac mass term in the Lagrangian. On the other hand, the Standard Model quantum numbers also forbid the appearance of a Majorana mass term for the neutrinos. In fact, the only dimensionful parameter in (61) is $\mu^2$, the parameter in the Higgs potential (63). Therefore, the mass of every particle in the model is expected to be proportional to $|\mu|$.

- Step 4: the next step of our program consists in choosing the parameter $\mu^2$ of the Higgs potential negative in order to trigger the phenomenon of spontaneous symmetry breaking and the Higgs mechanism. The minimum of the potential occurs at a point $v^2 = -\mu^2 / \lambda$. As we have explained earlier, we can choose the direction of the breaking to be along the real part of $\phi^0$.

- Step 5: translating the Higgs field by a real constant:

$$\Phi \rightarrow \Phi + \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \quad \lambda^2 = -\mu^2 / \lambda$$

(64)

transforms the Lagrangian and generates new terms, as it was explained in the previous section. Let us look at some of them:

(i) Fermion mass terms. Replacing $\phi^0$ by $v$ in the Yukawa term in (61) creates a mass term for the charged leptons, leaving the neutrinos massless.

$$m_e = \frac{1}{\sqrt{2}} G_e v \quad m_\mu = \frac{1}{\sqrt{2}} G_\mu v \quad m_\tau = \frac{1}{\sqrt{2}} G_\tau v$$

(65)

Since we have three arbitrary constants $G_i$, we can fit the three observed lepton masses. If we introduce right-handed neutrinos we can also fit whichever Dirac neutrino masses we wish.

(ii) Gauge boson mass terms. They come from the $|D_\mu \Phi|^2$ term in the Lagrangian. A straight substitution produces the following quadratic terms among the gauge boson fields:

$$\frac{1}{8} v^2 [g^2 (W_\mu^1 W_\mu^1 + W_\mu^2 W_\mu^2) + (g' B_\mu - g W_\mu^3)^2]$$

(66)

Defining the charged vector bosons as:
we obtain their masses:

$$m_W = \frac{v g}{2}$$

(68)

The neutral gauge bosons $B_\mu$ and $W^3_\mu$ have a $2 \times 2$ non-diagonal mass matrix. After diagonalisation, we define the mass eigenstates:

$$Z_\mu = \cos \theta W B_\mu - \sin \theta W W^3_\mu$$

$$A_\mu = \cos \theta W B_\mu + \sin \theta W W^3_\mu$$

with $\tan \theta_W = g'/g$. They correspond to the mass eigenvalues

$$m_Z = \frac{v(g^2 + g'^2)^{1/2}}{2} = \frac{m_W}{\cos \theta_W}$$

$$m_A = 0$$

(70)

As expected, one of the neutral gauge bosons is massless and will be identified with the photon. The Higgs mechanism breaks the original symmetry according to $SU(2) \times U(1) \rightarrow U(1)_{em}$ and $\theta_W$ is the angle between the original $U(1)$ and the one left unbroken. It is the parameter first introduced by S.L. Glashow, although it is often referred to as “Weinberg angle”.

(iii) Physical Higgs mass. Three out of the four real fields of the $\Phi$ doublet will be absorbed by the Higgs mechanism in order to allow for the three gauge bosons $W^\pm$ and $Z^0$ to acquire a mass. The fourth one, which corresponds to $(\phi^0\phi^0\dagger)^{1/2}$, remains physical. Its mass is given by the coefficient of the quadratic part of $V(\Phi)$ after the translation (64) and is equal to:

$$m_H = \sqrt{-2\mu^2} = \sqrt{2}\lambda v^2$$

(71)

In addition, we produce various coupling terms which we shall present, together with the hadronic ones, in the next section.

5.2 Extension to hadrons

Introducing the hadrons into the model presents some novel features. They are mainly due to the fact that the individual quark quantum numbers are not separately conserved. As regards to the second step, today there is a consensus regarding the choice of the “elementary” constituents of matter: Besides the six leptons, there are six quarks. They are fractionally charged and come each in three “colours”. The observed lepton-hadron universality property, tells us to use also doublets and singlets for the quarks. The first novel feature we mentioned above is that all quarks appear to have non-vanishing Dirac masses, so we must introduce both right-handed singlets for each family. A naïve assignment would be to write the analogue of Eqs. (56) and (57) as:

$$Q^i_L(x) = \frac{1}{2}(1 + \gamma_5) \left( \begin{array}{c} U^i(x) \\ D^i(x) \end{array} \right); \ U^i_R(x) ; \ D^i_R(x)$$

(72)
with the index \( i \) running over the three families as \( U^i = u, c, t \) and \( D^i = d, s, b \) for \( i = 1, 2, 3 \), respectively\(^8\). This assignment determines the \( SU(2) \) transformation properties of the quark fields. It also fixes their \( Y \) charges and, hence their \( U(1) \) properties. Using Eq. (55), we find

\[
Y(Q_L^i) = \frac{1}{3} \quad ; \quad Y(U^i_R) = \frac{4}{3} \quad ; \quad Y(D^i_R) = -\frac{2}{3}
\]  

(73)

The presence of the two right handed singlets has an important consequence. Even if we had only one family, we would have two distinct Yukawa terms between the quarks and the Higgs field of the form:

\[
\mathcal{L}_{Yuk} = G_d(\bar{Q}_L D_R \Phi + h.c.) + G_u(\bar{Q}_L U_R \Phi + h.c.)
\]  

(74)

\( \tilde{\Phi} \) is the doublet proportional to \( \tau_2 \Phi^* \). It has the same transformation properties under \( SU(2) \) as \( \Phi \), but the opposite \( Y \) charge.

If there were only one family, this would have been the end of the story. The hadron Lagrangian \( \mathcal{L}^{(1)}_h \) is the same as (61) with quark fields replacing leptons and the extra term of (74). The complication we alluded to before comes with the addition of more families. In this case the total Lagrangian is not just the sum over the family index. The physical reason is the non-conservation of the individual quark quantum numbers we mentioned previously. In writing (72), we implicitly assumed a particular pairing of the quarks in each family, \( u \) with \( d \), \( c \) with \( s \) and \( t \) with \( b \). In general, we could choose any basis in family space and, since we have two Yukawa terms, we will not be able to diagonalise both of them simultaneously. It follows that the most general Lagrangian will contain a matrix with non-diagonal terms which mix the families. By convention, we attribute it to a different choice of basis in the \( d - s - b \) space. It follows that the correct generalisation of the Yukawa Lagrangian (74) to many families is given by:

\[
\mathcal{L}_{Yuk} = \sum_{i,j} \left[ (\bar{Q}_L D_R^i \Phi + h.c.) \right] + \sum_i \left[ G_u^i (\bar{Q}_L U_R^i \Phi + h.c.) \right]
\]  

(75)

where the Yukawa coupling constant \( G_d \) has become a matrix in family space. After translation of the Higgs field, we shall produce masses for the up quarks given by \( m_u = G_u^{ij} v, m_c = G_c^{ij} v \) and \( m_t = G_t^{ij} v \), as well as a three-by-three mass matrix for the down quarks given by \( G_d^{ij} v \). As usually, we want to work in a field space where the masses are diagonal, so we change our initial \( d - s - b \) basis to bring \( G_d^{ij} \) into a diagonal form. This can be done through a three-by-three unitary matrix \( \tilde{D} = U^{ij} D^j \) such that \( U^i G_d^j U^j = \text{diag}(m_d, m_s, m_b) \). In the simplest example of only two families, it is easy to show that the most general such matrix, after using all freedom for field redefinitions and phase choices, is a real rotation:

\[
C = \begin{pmatrix}
\cos \theta & \sin \theta \\
-\sin \theta & \cos \theta
\end{pmatrix}
\]  

(76)

with \( \theta \) being our familiar Cabibbo angle. For three families an easy counting shows that the matrix has three angles, the three Euler angles, and an arbitrary phase. It is traditionally written in the form:

\[
KM = \begin{pmatrix}
c_1 s_1 c_3 & s_1 c_3 & s_1 s_3 \\
-s_1 c_3 & c_1 c_2 c_3 - s_2 s_3 e^{i\delta} & c_1 c_2 s_3 + s_2 c_3 e^{i\delta} \\
-s_1 s_2 & c_1 s_2 c_3 + c_2 s_3 e^{i\delta} & c_1 s_2 s_3 - c_2 c_3 e^{i\delta}
\end{pmatrix}
\]  

(77)

with the notation \( c_k = \cos \theta_k \) and \( s_k = \sin \theta_k, k = 1, 2, 3 \). The novel feature is the possibility of introducing the phase \( \delta \). This means that a six-quark model has a natural source of \( CP \), or \( T \), violation, while a four-quark model does not.

\(^8\)An additional index \( a \), running also through 1, 2 and 3 and denoting the colour, is understood.
The total Lagrangian density, before the translation of the Higgs field, is now:

\[
\mathcal{L} = -\frac{1}{4} \vec{W}_{\mu\nu} \cdot \vec{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} + |D_\mu \Phi|^2 - V(\Phi)
\]

\[+ \sum_{i=1}^{3} \left[ \vec{\Psi}^i_L i\gamma^\mu \partial_\mu \vec{\Psi}^i_L + \vec{R}_i i\gamma^\mu \partial_\mu \vec{R}_i - G_i (\vec{\Psi}^i_L R_i \Phi + h.c.) \right] \]

\[+ \sum_{i,j=1}^{3} \left[ (\vec{Q}^i_L G^{ij}_R \vec{D}^j_R \Phi + h.c.) \right] \]

(78)

The covariant derivatives on the quark fields are given by:

\[
D_\mu Q^i_L = \left( \partial_\mu - ig \frac{2}{3} \vec{W}_\mu - ig' \frac{1}{6} B_\mu \right) Q^i_L
\]

\[
D_\mu U^i_R = \left( \partial_\mu - ig' \frac{2}{3} B_\mu \right) U^i_R
\]

\[
D_\mu D^i_R = \left( \partial_\mu + ig' \frac{2}{3} B_\mu \right) D^i_R
\]

(79)

The classical Lagrangian (78) contains seventeen arbitrary real parameters. They are:

- the two gauge coupling constants \( g \) and \( g' \);
- the two parameters of the Higgs potential \( \lambda \) and \( \mu^2 \);
- three Yukawa coupling constants for the three lepton families, \( G_{e,\mu,\tau} \);
- six Yukawa coupling constants for the three quark families, \( G^{u,c,t}_u \) and \( G^{d,s,b}_d \);
- four parameters of the \( KM \) matrix, the three angles and the phase \( \delta \).

A final remark: Fifteen out of these seventeen parameters are directly connected with the Higgs sector.

Translating the Higgs field by Eq. (64) and diagonalising the resulting down quark mass matrix produces the mass terms for fermions and bosons which we introduced before as well as several coupling terms. We shall write here the ones which involve the physical fields\(^9\).

(i) The gauge boson fermion couplings. They are the ones which generate the known weak and electromagnetic interactions. \( A_\mu \) is coupled to the charged fermions through the usual electromagnetic current.

\[
\frac{gg'}{(g^2 + g'^2)^{1/2}} \left[ e^\gamma_\mu e + \sum_{a=1}^{3} \left( \frac{2}{3} \bar{u}^a \gamma_\mu u^a - \frac{1}{3} \bar{d}^a \gamma_\mu d^a \right) \right] A_\mu
\]

(80)

where the dots stand for the contribution of the other two families \( e \rightarrow \mu, \tau \), \( u \rightarrow c, t \) and \( d \rightarrow s, b \) and the summation over \( a \) extends over the three colours. Equation (80) shows that the electric charge \( e \) is given, in terms of \( g \) and \( g' \) by

\[
e = \frac{gg'}{(g^2 + g'^2)^{1/2}} = g \sin \theta_W = g' \cos \theta_W
\]

(81)

\(^9\text{We know from quantum electrodynamics that, in order to determine the Feynman rules of a gauge theory, one must first decide on a choice of gauge. For Yang-Mills theories this step introduces new fields called Faddeev-Popov ghosts. This point is explained in every standard text book on quantum field theory, but we have not discussed it in these lectures.}\)
Similarly, the couplings of the charged $W$'s to the weak current are:

$$\frac{g}{2\sqrt{2}} \left( \bar{\nu}_e \gamma^\mu (1 + \gamma^5)e^\mu + \sum_{a=1}^{3} \bar{u}_a \gamma^\mu (1 + \gamma^5) \gamma^\nu d^\mu_{KM} \right) + \ldots = \partial^\mu W^\pm_\nu - \partial^\nu W^\pm_\mu$$

Combining all these relations, we can determine the experimental value of the parameter $v$, the vacuum expectation value of the Higgs field. We find $v \sim 246 \text{GeV}$.

As expected, only left-handed fermions participate. $d_{KM}$ is the linear combination of $d - s - b$ given by the KM matrix (77). By diagonalising the down quark mass matrix we introduced the off-diagonal terms into the hadron current. When considering processes, like nuclear $\beta$-decay, or $\mu$-decay, where the momentum transfer is very small compared to the $W$ mass, the $W$ propagator can be approximated by $m_W^{-2}$ and the effective Fermi coupling constant is given by:

$$G = \frac{g^2}{8m_W^2} = \frac{1}{2v^2}$$

Contrary to the charged weak current (82), the $Z^0$-fermion couplings involve both left- and right-handed fermions:

$$-\frac{e}{2} \sin \theta_W \cos \theta_W \left[ \bar{\nu}_L \gamma^\mu \nu_L + (\sin^2 \theta_W - \cos^2 \theta_W) \bar{e}_L \gamma^\mu e_L 
+ 2\sin^2 \theta_W \bar{e}_R \gamma^\mu e_R + \ldots \right] Z_\mu$$

$$= \frac{e}{2} \sum_{a=1}^{3} \left[ \left( \frac{1}{3} \tan \theta_W - \cot \theta_W \right) \bar{u}_a^L \gamma^\mu u^L_a + \left( \frac{1}{3} \tan \theta_W + \cot \theta_W \right) \bar{d}_a^L \gamma^\mu d^L_a 
+ \frac{2}{3} \tan \theta_W \left( 2\bar{u}_a^R \gamma^\mu u^R_a - \bar{d}_a^R \gamma^\mu d^R_a \right) + \ldots \right] Z_\mu$$

Again, the summation is over the colour indices and the dots stand for the contribution of the other two families. We verify in this formula the property of the weak neutral current to be diagonal in the quark flavour space. Another interesting property is that the axial part of the neutral current is proportional to $[\bar{u} \gamma^\mu \gamma^5 u - \bar{d} \gamma^\mu \gamma^5 d]$. This particular form of the coupling is important for the phenomenological applications, such as the induced parity violating effects in atoms and nuclei.

(ii) The gauge boson self-couplings. One of the characteristic features of Yang-Mills theories is the particular form of the self couplings among the gauge bosons. They come from the square of the non-Abelian curvature in the Lagrangian, which, in our case, is the term $-\frac{1}{4} W^\lambda_{\mu \nu} \cdot W^\lambda_{\mu \nu}$. Expressed in terms of the physical fields, this term gives:

$$-ig(\sin \theta_W A^\mu - \cos \theta_W Z^\mu)(W^\nu W^\mu_{\mu \nu} - W^\mu W^\nu_{\mu \nu})$$
$$-ig(\sin \theta_W F^{\mu \nu} - \cos \theta_W Z^{\mu \nu})W^\nu_{\mu \nu} W^\mu_{\nu \mu}$$
$$\frac{g^2}{2}(\sin^2 \theta_W A^\mu - \cos \theta_W Z^\mu)^2 W^\nu_{\mu \nu} W^\nu_{\mu \nu}$$
$$+ g^2(\sin \theta_W A^\mu - \cos \theta_W Z^\mu)(\sin \theta_W A^\nu - \cos \theta_W Z^\nu)W^\mu_{\nu \mu} W^\mu_{\nu \nu}$$
$$- \frac{g^2}{2} (W^\mu W^\nu - W^\mu W^\nu)^2$$

where we have used the following notation: $F_{\mu \nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$, $W^\pm_{\mu \nu} = \partial_\mu W^\pm_{\nu \mu} - \partial_\nu W^\pm_{\mu \nu}$ and $Z_{\mu \nu} = \partial_\mu Z_\nu - \partial_\nu Z_\mu$ with $g \sin \theta_W = e$. Let us concentrate on the photon-$W^+ W^-$ couplings. If we
forget, for the moment, about the $SU(2)$ gauge invariance, we can use different coupling constants for the two trilinear couplings in (86), say $e$ for the first and $e\kappa$ for the second. For a charged, massive $W$, the magnetic moment $\mu$ and the quadrupole moment $Q$ are given by:

$$
\mu = \frac{(1 + \kappa)e}{2m_W} \quad Q = -\frac{e\kappa}{m_W^2}
$$

(87)

Looking at (86), we see that $\kappa = 1$. Therefore, $SU(2)$ gauge invariance gives very specific predictions concerning the electromagnetic parameters of the charged vector bosons. The gyromagnetic ratio equals two and the quadrupole moment equals $-em_W^2$.

(iii) The scalar Higgs fermion couplings. They are given by the Yukawa terms in (61). The same couplings generate the fermion masses through spontaneous symmetry breaking. It follows that the physical Higgs scalar couples to quarks and leptons with strength proportional to the fermion mass. Therefore the prediction is that it will decay predominantly to the heaviest possible fermion compatible with phase space. This property provides a typical signature for Higgs identification.

(iv) The scalar Higgs gauge boson couplings. They come from the covariant derivative term $|D_\mu \Phi|^2$ in the Lagrangian. If we call $\phi$ the field of the physical neutral Higgs, we find:

$$\frac{1}{4}(v + \phi)^2 \left[ g^2 W^+ W^- + (g^2 + g'^2)Z^\mu Z^\mu \right]$$

(88)

This gives a direct coupling $\phi - W^+ - W^-$, as well as $\phi - Z - Z$, which has been very useful in the Higgs searches.

(v) The scalar Higgs self couplings. They are proportional to $\lambda(v + \phi)^4$. Equations (71) and (83) show that $\lambda = Gm_H^2/\sqrt{2}$, so, in the tree approximation, this coupling is related to the Higgs mass. It could provide a test of the Standard Model Higgs, but it will not be easy to measure. On the other hand this relation shows that, if the physical Higgs is very heavy, it is also strongly interacting and this sector of the model becomes non-perturbative.

The five step program is now complete for both leptons and quarks. The seventeen parameters of the model have all been determined by experiment. Although the number of arbitrary parameters seems very large, we should not forget that they are all mass and coupling parameters, like the electron mass and the fine structure constant of quantum electrodynamics. The reason we have more of them is that the Standard Model describes in a unified framework a much larger number of particles and interactions.

6 The Standard Model and experiment

Our confidence in this model is amply justified on the basis of its ability to accurately describe the bulk of our present-day data and, especially, of its enormous success in predicting new phenomena. Let us mention a few of them. We shall follow the historical order.

- The discovery of weak neutral currents by Gargamelle in 1972

$$\nu_\mu + e^- \to \nu_\mu + e^- ; \quad \nu_\mu + N \to \nu_\mu + X$$

Both, their strength and their properties were predicted by the Model.

- The discovery of charmed particles at SLAC in 1974

Their presence was essential to ensure the absence of strangeness changing neutral currents, ex. $K^0 \to \mu^+ + \mu^-$

Their characteristic property is to decay predominantly in strange particles.

- A necessary condition for the consistency of the Model is that $\sum_i Q_i = 0$ inside each family.
When the $\tau$ lepton was discovered this implied a prediction for the existence of the $b$ and $t$ quarks with the right electric charges.

- The discovery of the $W$ and $Z$ bosons at CERN in 1983 with the masses predicted by the theory.

The characteristic relation of the Standard Model with an isodoublet Higgs mechanism $m_Z = m_W / \cos \theta_W$ has been checked with very high accuracy (including radiative corrections).

- The $t$-quark was seen at LEP through its effects in radiative corrections before its actual discovery at Fermilab.

- The vector boson self-couplings, $\gamma-W^+-W^-$ and $Z^0-W^+-W^-$ have been measured at LEP and confirm the Yang-Mills predictions given in equation (87).

- The recent discovery of a new boson which could be the Higgs particle of the Standard Model is the last of this impressive series of successes.

All these discoveries should not make us forget that the Standard Model has been equally successful in fitting a large number of experimental results. You have all seen the global fit given in Fig. 5. The conclusion is obvious: The Standard Model has been enormously successful.

Although in these lectures we did not discuss quantum chromodynamics, the gauge theory of strong interactions, the computations whose results are presented in Fig. 5, take into account the radiative corrections induced by virtual gluon exchanges. The fundamental property of quantum chromodynamics, the one which allows for perturbation theory calculations, is the property of asymptotic freedom, the particular dependence of the effective coupling strength on the energy scale. This is presented in Fig. 6. The green region shows the theoretical prediction based on QCD calculations, including the theoretical uncertainties. We see that the agreement with the experimentally measured values of the effective strong interaction coupling constant $\alpha_s$ is truly remarkable. Notice also that this agreement extends to rather low values of $Q$ of the order of 1-2 GeV, where $\alpha_s$ equals approximately 1/3.

This brings us to our next point, namely that all this success is in fact a success of renormalised
The extreme accuracy of the experimental measurements, mainly at LEP, but also at FermiLab and elsewhere, allow, for the first time to make a detailed comparison between theory and experiment including the purely weak interaction radiative corrections.

In Fig. 7 we show the comparison between theory and experiment for two quantities, $\epsilon_1$ and $\epsilon_3$, defined in equations (89) and (90), respectively:

$$\epsilon_1 = \frac{3G_F m_t^2}{8\sqrt{2}\pi^2} - \frac{3G_F m_W^2}{4\sqrt{2}\pi^2} \tan^2 \theta_W \ln \frac{m_H}{m_Z} + ...$$

$$\epsilon_3 = \frac{G_F m_W^2}{12\sqrt{2}\pi^2} \ln \frac{m_H}{m_Z} - \frac{G_F m_W^2}{6\sqrt{2}\pi^2} \ln \frac{m_t}{m_Z} + ...$$

They are defined with the following properties: (i) They include the strong and electromagnetic radiative corrections and (ii), they vanish in the Born approximation for the weak interactions. So, they measure the purely weak interaction radiative corrections. The figure is based on a fit which is rather old and does not include the latest data but, nevertheless, it shows that, in order to obtain agreement with the data, one must include these corrections. Weak interactions are no more a simple phenomenological model, but have become a precision theory.

The moral of the story is that the perturbation expansion of the Standard Model is reliable as long as all coupling constants remain small. The only coupling which does become large in some kinematical regions is $\alpha_s$ which grows at small energy scales, as shown in Fig. 6. In this region we know that a hadronisation process occurs and perturbation theory breaks down. We conclude that at high energies perturbation theory is expected to be reliable unless there are new strong interactions.

This brings us to our last point, namely that this very success shows also that the Standard Model cannot be a complete theory, in other words there must be new Physics beyond the Standard Model. The argument is simple and it is based on a straightforward application of perturbation theory with an additional assumption which we shall explain presently.

---

10Updated electroweak fits can be found at the following URL: http://gfitter.desy.de.
Fig. 7: Comparison between theory and experiment for two quantities sensitive to weak interaction radiative corrections.

We assume that the Standard Model is correct up to a certain scale $\Lambda$. The precise value of $\Lambda$ does not matter, provided it is larger than any energy scale reached so far\textsuperscript{11}.

A quantum field theory is defined through a functional integral over all classical field configurations, the Feynman path integral. By a Fourier transformation we can express it as an integral over the fields defined in momentum space. Following K. Wilson, let us split this integral in two parts: the high energy part with modes above $\Lambda$ and the low energy part with the modes below $\Lambda$. Let us imagine that we perform the high energy part. The result will be an effective theory expressed in terms of the low energy modes of the fields. We do not know how to perform this integration explicitly, so we cannot write down the correct low energy theory, but the most general form will be a series of operators made out of powers of the fields and their derivatives. Since integrating over the heavy modes does not break any of the symmetries of the initial Lagrangian, only operators allowed by the symmetries will appear. Wilson remarked that, when $\Lambda$ is large compared to the mass parameters of the theory, we can determine the leading contributions by simple dimensional analysis\textsuperscript{12}. We distinguish three kinds of operators, according to their canonical dimension:

- those with dimension larger than four. Dimensional analysis shows that they will come with a coefficient proportional to inverse powers of $\Lambda$, so, by choosing the scale large enough, we can make their contribution arbitrarily small. We shall call them \textit{irrelevant operators};
- those with dimension equal to four. They are the ones which appeared already in the original Lagrangian. Their coefficient will be independent of $\Lambda$, up to logarithmic corrections which we ignore. We shall call them \textit{marginal operators};
- finally we have the operators with dimension smaller than four. In the Standard Model there

\textsuperscript{11}The scale $\Lambda$ should not be confused with a cut-off one often introduces when computing Feynman diagrams. This cut-off disappears after renormalisation is performed. Here $\Lambda$ is a physical scale which indicates how far the theory can be trusted.

\textsuperscript{12}There are some additional technical assumptions concerning the dimensions of the fields, but they are satisfied in perturbation theory.
is only one such operator, the square of the Higgs field $\Phi^2$ which has dimension equal to two\textsuperscript{13}. This operator will appear with a coefficient proportional to $\Lambda^2$, which means that its contribution will grow quadratically with $\Lambda$. We shall call it 	extit{relevant operator}. It will give an effective mass to the scalar field proportional to the square of whichever scale we can think of. This problem was first identified in the framework of Grand Unified Theories and is known since as the 	extit{hierarchy problem}. Let me emphasise here that this does not mean that the mass of the scalar particle will be necessarily equal to $\Lambda$. The Standard Model is a renormalisable theory and the mass is fixed by a renormalisation condition to its physical value. It only means that this condition should be adjusted to arbitrary precision order by order in perturbation theory. It is this extreme sensitivity to high scales, known as the 	extit{fine tuning problem}, which is considered unacceptable for a fundamental theory.

Let us summarise: The great success of the Standard Model tells us that renormalised perturbation theory is reliable in the absence of strong interactions. The same perturbation theory shows the need of a fine tuning for the mass of the scalar particle. If we do not accept the latter, we have the following two options:

- perturbation theory breaks down at some scale $\Lambda$. We can imagine several reasons for such a breakdown to occur. The simplest is the appearance of new strong interactions. The so called 	extit{Technicolor} models, in which the role of the Higgs field is played by a bound state of new strongly coupled fermions, were in this class. More exotic possibilities include the appearance of new, compact space dimensions with compactification length $\sim \Lambda^{-1}$;

- perturbation theory is still valid but the numerical coefficient of the $\Lambda^2$ term which multiplies the $\Phi^2$ operator vanishes to all orders of perturbation theory. For this to happen we must modify the Standard Model introducing appropriate new particles. Supersymmetry is the only systematic way we know to achieve this goal.

7 Conclusions

In these lectures we saw the fundamental role of Geometry in the Dynamics of the forces among the elementary particles. It was the understanding of this role which revolutionised our way of thinking and led to the construction of the Standard Model. It incorporates the ideas of gauge theories, as well as those of spontaneous symmetry breaking. Its agreement with experiment is spectacular. It fits all data known today. However, unless one is willing to accept a fine tuning with arbitrary precision, one should conclude that New Physics will appear beyond a scale $\Lambda$. The precise value of $\Lambda$ cannot be computed, but the amount of fine tuning grows quadratically with it, so it cannot be too large. Hopefully, it will be within reach of the LHC.

Acknowledgments

I wish to thank all the participants to the 2012 CERN Summer School who, by their questions and remarks, helped me in sharpening the arguments presented in these notes and, in particular, Dr Christophe Grojean for a critical reading of the manuscript.

\textsuperscript{13}One could think of the square of a fermion operator $\bar{\Psi}\Psi$, whose dimension is equal to three, but it is not allowed by the chiral symmetry of the model.
A brief Introduction to Modern Amplitude Methods

L. J. Dixon
SLAC National Accelerator Laboratory, Stanford University, Stanford, USA

Abstract
I provide a basic introduction to modern helicity amplitude methods, including color organization, the spinor helicity formalism, and factorization properties. I also describe the BCFW (on-shell) recursion relation at tree level, and explain how similar ideas — unitarity and on-shell methods — work at the loop level. These notes are based on lectures delivered at the 2012 CERN Summer School and at TASI 2013.

1 Introduction
Scattering amplitudes are at the heart of high energy physics. They lie at the intersection between quantum field theory and collider experiments. Currently we are in the hadron collider era, which began at the Tevatron and has now moved to the Large Hadron Collider (LHC). Hadron colliders are broadband machines capable of great discoveries, such as the Higgs boson [1], but there are also huge Standard Model backgrounds to many potential signals. If we are to discover new physics (besides the Higgs boson) at the LHC, we will need to understand the old physics of the Standard Model at an exquisitely precise level. QCD dominates collisions at the LHC, and the largest theoretical uncertainties for most processes are due to our limited knowledge of higher order terms in perturbative QCD.

Many theorists have been working to improve this situation. Some have been computing the next-to-leading order (NLO) QCD corrections to complex collider processes that were previously only known at leading order (LO). LO uncertainties are often of order one, while NLO uncertainties can be in the 10–20% range, depending on the process. Others have been computing the next-to-next-to-leading order (NNLO) corrections to benchmark processes that are only known at NLO; most NNLO predictions have uncertainties in the range of 1–5%, allowing precise experimental measurements to be interpreted with similar theoretical precision. Higher-order computations have a number of technical ingredients, but they all require loop amplitudes, one-loop for NLO, and both one- and two-loop for NNLO, as well as tree amplitudes of higher multiplicity.

The usual textbook methods for computing an unpolarized cross section involve squaring the scattering amplitude at the beginning, then summing analytically over the spins of external states, and transforming the result into an expression that only involves momentum invariants (Mandelstam variables) and masses. For complex processes, this approach is usually infeasible. If there are $N$ Feynman diagrams for an amplitude, then there are $N^2$ terms in the square of the amplitude. It is much better to calculate the $N$ terms in the amplitude, as a complex number, and then compute the cross section by squaring that number. This approach of directly computing the amplitude benefits greatly from the fact that many amplitudes are much simpler than one might expect from the number of Feynman diagrams contributing to them.

In order to compute the amplitude directly, one has to pick a basis for the polarization states of the external particles. At collider energies, most of these particles are effectively massless: the light quarks and gluons, photons, and the charged leptons and neutrinos (decay products of $W$ and $Z$ bosons). Massless fermions have the property that their chirality and helicity coincide, and their chirality is preserved by the gauge interactions. Therefore the helicity basis is clearly an optimal one for massless fermions, because many matrix elements (the helicity-flip ones) will always vanish.

Around three decades ago, it was realized that the helicity basis was extremely useful for massless vector bosons as well [2]. Many tree-level amplitudes were found to vanish in this basis as well (which
could be explained by a secret supersymmetry obeyed by tree amplitudes [3, 4]). Also, the nonvanishing amplitudes were found to possess a hierarchy of simplicity, depending on how much they violated helicity “conservation”. For example, a simple one-term expression for the tree amplitudes for scattering an arbitrary number of gluons with maximal helicity violation (MHV) was found by Parke and Taylor in 1986 [5], and proven recursively by Berends and Giele shortly thereafter [6].

As the first loop computations were performed for gluon scattering in the helicity basis [7, 8], it became apparent that (relative) simplicity of amplitudes could extend to the loop level. One way to maintain the simplicity is to use unitarity [9] to determine loop amplitudes by using tree amplitudes as input. These methods have been refined enormously over the years, and automated in order to handle very complicated processes. They now form an important part of the arsenal for theorists providing NLO results for LHC experiments. Many of the methods are now being further refined and extended to the two-loop level, and within a few years we may see a similar NNLO arsenal come to full fruition.

Besides QCD, unitarity-based methods have also found widespread application to scattering amplitudes for more formal theories, such as $\mathcal{N} = 4$ super-Yang-Mills theory and $\mathcal{N} = 8$ supergravity, just to mention a couple of examples. The more supersymmetry, the greater the simplicity of the amplitudes, allowing analytical results to be obtained for many multi-loop amplitudes (at least before integrating over the loop momentum). These results have helped to expose new symmetries, which have in turn led to other powerful methods for computing in these special theories.

The purpose of these lecture notes is to provide a brief and basic introduction to modern amplitude methods. They are intended for someone who has taken a first course in quantum field theory, but who has never studied these methods before. For someone who wants to go on further and perform research using such methods in either QCD or more formal areas, these notes will be far from sufficient. Fortunately, there are much more thorough reviews available. In particular, methods for one-loop QCD amplitudes have been reviewed in refs. [10–13]. Also, a very recent and comprehensive article [14] covers much of the material covered here, plus a great deal more, particularly in the direction of methods for multi-loop amplitudes in more formal theories. There are also reviews of basic tree-level organization and properties [15–17] and of one-loop unitarity [18]. Other reviews emphasize $\mathcal{N} = 4$ super-Yang-Mills theory [19, 20].

These notes are organized as follows. In Section 2 we describe trace-based color decompositions for QCD amplitudes. In Section 3 we review the spinor helicity formalism, and apply it to the computation of some simple four- and five-point tree amplitudes. In Section 4 we use these results to illustrate the universal soft and collinear factorization of gauge theory amplitudes. We also introduce the Parke-Taylor amplitudes, and discuss the utility of spinor variables for describing collinear limits and massless three-point kinematics. In Section 5 we explain the BCFW (on-shell) recursion relation for tree amplitudes, and apply it to the Parke-Taylor amplitudes, as well as to a next-to-MHV example. Section 6 discusses the application of generalized unitarity to one-loop amplitudes, and in Section 7 we conclude.

2 Color decompositions

In this section we explain how to organize the color degrees of freedom in QCD amplitudes, in order to separate out pieces that have simpler analytic properties. Those pieces have various names in the literature, such as color-ordered amplitudes, dual amplitudes, primitive amplitudes and partial amplitudes. (There is a distinction between primitive amplitudes and partial amplitudes at the loop level, but not at tree level, at least not unless there are multiple fermion lines.)

The basic idea [15, 16, 21, 22] is to reorganize the color degrees of freedom of QCD, in order to eliminate the Lie algebra structure constants $f^{abc}$ found in the Feynman rules, in favor of the generator matrices $T^a$ in the fundamental representation of $SU(N_c)$. Although the gauge group of QCD is $SU(3)$, it requires no extra effort to generalize it to $SU(N_c)$, and one can often gain insight by making the dependence on $N_c$ explicit. Sometimes it is also advantageous (especially computationally) to consider
the limit of a large number of colors, $N_c \to \infty$.

Gluons in an $SU(N_c)$ gauge theory carry an adjoint color index $a = 1, 2, \ldots, N_c^2 - 1$, while quarks and antiquarks carry an $N_c$ or $\overline{N}_c$ index, $i, \bar{i} = 1, 2, \ldots, N_c$. The generators of $SU(N_c)$ in the fundamental representation are traceless hermitian $N_c \times N_c$ matrices, $(T^a)^i_{\ j}$. For computing color-ordered helicity amplitudes, it’s conventional to normalize them according to $\text{Tr}(T^a T^b) = \delta^{ab}$ in order to avoid a proliferation of $\sqrt{2}$’s in the amplitudes.

Each Feynman diagram in QCD contains a factor of $(T^a)^i_{\ j}$ for each gluon-quark-anti-quark vertex, a group theory structure constant $f^{abc}$ for each pure gluon three-point vertex, and contracted pairs of structure constants $f^{abc} f^{cde}$ for each pure gluon four-vertex. The structure constants are defined by the commutator

$$[T^a, T^b] = i\sqrt{2} f^{abc} T^c. \quad (1)$$

The internal gluon and quark propagators contract indices together with factors of $\delta_{ab}, \delta^i_{\ j}$. We want to identify all possible color factors for the diagrams, and sort the contributions into gauge-invariant subsets with simpler analytic properties than the full amplitude.

To do this, we first eliminate all the structure constants $f^{abc}$ in favor of the generators $T^a$, using

$$\tilde{f}^{abc} \equiv i\sqrt{2} f^{abc} = \text{Tr}(T^a T^b T^c) - \text{Tr}(T^a T^c T^b), \quad (2)$$

which follows from the definition (1) of the structure constants. This identity is represented graphically in fig. 1(a), in which curly lines are in the adjoint representation and lines with arrows are in the fundamental representation. After this step, every color factor for a multi-gluon amplitude is a product of some number of traces. Many traces share $T^a$’s with contracted indices, of the form $\text{Tr}(\ldots T^a \ldots) \text{Tr}(\ldots T^a \ldots) \ldots \text{Tr}(\ldots)$. If external quarks are present, then in addition to the traces there will be some strings of $T^a$’s terminated by fundamental indices, of the form $(T^{a_1} \ldots T^{a_m})^i_{\ j_2}$. In order to reduce the number of traces and strings we can apply the $SU(N_c)$ Fierz identity,

$$(T^a)^{\bar{i}_1}_{\ i_1} (T^a)^{\bar{i}_2}_{\ i_2} = \delta^{\bar{i}_2}_{\ i_1} \delta^{\bar{i}_1}_{\ i_2} - \frac{1}{N_c} \delta^{\bar{i}_1}_{\ i_1} \delta^{\bar{i}_2}_{\ i_2}, \quad (3)$$

where the sum over $a$ is implicit. This identity is illustrated graphically in Fig.1(b).

Equation (3) is just the statement that the $SU(N_c)$ generators $T^a$ form the complete set of traceless hermitian $N_c \times N_c$ matrices. The $-1/N_c$ term implements the tracelessness condition. (To see this,
contract both sides of Eq. (3) with $\delta j^i_1$.) It is often convenient to consider also $U(N_c) = SU(N_c) \times U(1)$ gauge theory. The additional $U(1)$ generator is proportional to the identity matrix,

$$ (T^{a U(1)})^j_i = \frac{1}{\sqrt{N_c}} \delta^j_i; \quad (4) $$

when this generator is included in the sum over $a$ in Eq. (3), the corresponding $U(N_c)$ result is Eq. (3) without the $-1/N_c$ term. The auxiliary $U(1)$ gauge field is often referred to as a photon. It is colorless, commuting with $SU(N_c)$, with vanishing structure constants $f^{a U(1) bc} = 0$ for all $b, c$. Therefore it does not couple directly to gluons, although quarks carry charge under it. Real photon amplitudes can be obtained using this generator, after replacing factors of the strong coupling $g$ with the QED coupling $\sqrt{2}e$.

The color algebra can easily be carried out graphically [23], as illustrated in Fig. 2. Starting with any given Feynman diagram, one interprets it as just the color factor for the full diagram, after expanding the four-gluon vertices into two three-gluon vertices. Then one makes the two substitutions, Eqs. (2) and (3), which are represented diagrammatically in Fig. 1. In Fig. 2 we use these steps to simplify a sample diagram for five-gluon scattering at tree level. Inserting the rule Fig. 1(a) in the three vertices leads to $2^3 = 8$ terms, of which two are shown in the first line. The $SU(N_c)$ Fierz identity takes the traces of products of three $T^a$’s, and systematically combines them into a single trace, $\text{Tr}(T^{a_1}T^{a_2}T^{a_3}T^{a_4}T^{a_5})$, plus all possible permutations, as shown in the second line of the figure. Notice that the $-1/N_c$ term in Eq. (3) and Fig. 1(b) does not contribute here, because the photon does not couple to gluons; that is, $f^{abI} = 0$ when $I$ is the $U(1)$ generator. (The $-1/N_c$ term only has to be retained when a gluon can couple to a fermion line at both ends.)

From Fig. 2 it is clear that any tree diagram for $n$-gluon scattering can be reduced to a sum of “single trace” terms, in which the generators $T^{a_i}$ corresponding to the external gluons have different cyclic orderings. The color decomposition of the the $n$-gluon tree amplitude [21] is,

$$ A_n^{\text{tree}}(\{k_i, \lambda_i, a_i\}) = g^{n-2} \sum_{\sigma \in S_n/Z_n} \text{Tr}(T^{a_{\sigma(1)}} \ldots T^{a_{\sigma(n)}}) A_n^{\text{tree}}(\sigma(1^{\lambda_1}), \ldots, \sigma(n^{\lambda_n})). \quad (5) $$

Here $g$ is the gauge coupling ($g^2/(4\pi) = \alpha_s$), $k_i, \lambda_i$ are the gluon momenta and helicities, and $A_n^{\text{tree}}(1^{\lambda_1}, \ldots, n^{\lambda_n})$ are the partial amplitudes, which contain all the kinematic information. $S_n$ is the set of all permutations of $n$ objects, while $Z_n$ is the subset of cyclic permutations, which preserves the trace; one sums over the set $S_n/Z_n$ in order to sweep out all cyclically-inequivalent orderings in the trace. We write the helicity label for each particle, $\lambda_i = \pm$, as a superscript.
The real work is in calculating the independent partial amplitudes $A_{\text{tree}}^n$. However, they are simpler than the full amplitude because they are color-ordered: they only receive contributions from diagrams with a particular cyclic ordering of the gluons. This feature reduces the number of singularities they can contain. Tree amplitudes contain factorization poles, when a single intermediate state goes on its mass shell in the middle of the diagram. The momentum of the intermediate state is the sum of a number of the external momenta. In the color-ordered partial amplitudes, those momenta must be cyclically adjacent in order to produce a pole. For example, the five-point partial amplitudes $A^5_5(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ can only have poles in $s_{12}, s_{23}, s_{34}, s_{45},$ and $s_{51}$, and not in $s_{13}, s_{24}, s_{35}, s_{41},$ or $s_{52}$, where $s_{ij} \equiv (k_i + k_j)^2$. Similarly, at the loop level, the only channels made out of sums of cyclically adjacent momenta will have unitarity cuts (as well as factorization poles). The number of cyclically-adjacent momentum channels grows much more slowly than the total number of channels, as the number of legs increases. Later we will use factorization properties to construct tree amplitudes, so defining partial amplitudes with a minimal number of factorization channels will simplify the construction.

Although we have mainly considered the pure-gluon case, color decompositions can be found for generic QCD amplitudes. Another simple case is the set of tree amplitudes $\bar{q}qgg \ldots g$ with two external quarks, which can be reduced to single strings of $T^a$ matrices,

$$A_{\text{tree}}^n = g^{n-2} \sum_{\sigma \in S_{n-2}} (T^{a_{\sigma(3)}} \ldots T^{a_{\sigma(n)}})_{ij} \bar{A}_{\text{tree}}^n(1^{\lambda_1}, 2^{\lambda_2}, \sigma(3^{\lambda_3}), \ldots, \sigma(n^{\lambda_n})), \quad (6)$$

where numbers without subscripts refer to gluons. Color decompositions for tree amplitudes with more than two external quarks can be found in Ref. [15].

The same ideas also work at the loop level [24]. For example, at one loop, the same graphical analysis leads to a color decomposition for pure-gluon amplitudes which contains two types of terms:

- single-trace terms, of the form $N_c \text{ Tr}(T^{a_1} \ldots T^{a_n})$ plus permutations, which contain an extra factor of $N_c$, and dominate for large $N_c$, and
- double-trace terms, of the form $\text{ Tr}(T^{a_1} \ldots T^{a_m}) \text{ Tr}(T^{a_{m+1}} \ldots T^{a_n})$ plus permutations, whose contribution to the color-summed cross section is suppressed by at least a factor of $1/N_c^2$ with respect to the leading-color terms.

Quark loops lead to contributions of the first type, but with an over all factor of the number of light quark flavors, $n_f$, replacing the factor of $N_c$.

After we have computed all of the partial amplitudes, the parton model requires us to construct the squared amplitude, averaged over the colors of the initial-state partons, and summed over the final-state colors. Using the above color decompositions, and applying Fierz identities again, this color-summed cross section can be expressed in terms of the partial amplitudes. The color factors that appear can be computed graphically. Take a single trace structure of the type shown in Fig. 2, and glue the $n$ gluon lines to a second trace structure from the conjugate amplitude, which may have a relative permutation. Then apply the Fierz identity in Fig. 1(b) to remove the gluon lines and reduce the resulting “vacuum” color graph to powers of $N_c$. (A closed loop for an arrowed line gives a factor of $\text{ Tr}(1) = N_c$.)

In this way you can show that the color-summed cross section for $n$-gluon scattering,

$$d\sigma_{\text{tree}}^n \propto \sum_{a_i=1}^{N_c^2-1} |A_{\text{tree}}^n(\{k_i, a_i\})|^2, \quad (7)$$

takes the form,

$$d\sigma_{\text{tree}}^n \propto N_c^n \left\{ \sum_{\sigma \in S_n/Z_n} |A_{\text{tree}}^n(\sigma(1), \sigma(2), \ldots, \sigma(n))|^2 + O(1/N_c^2) \right\}. \quad (8)$$
In other words, the leading-color contributions come from gluing together two trace structures with no relative permutation, which gives rise to a planar vacuum color graph. Any relative permutation leads to a nonplanar graph, and its evaluation results in at least two fewer powers of $N_c$. Of course these subleading-color terms can be worked out straightforwardly as well. Another way of stating Eq. (8) is that, up to $1/N_c^2$-suppressed terms, the differential cross section can be written as a sum of positive terms, each of which has a definite color flow. This description is important for the development of parton showers, which exploit the pattern of radiating additional soft gluons from these color-ordered pieces of the cross section.

3 The spinor helicity formalism

3.1 Spinor variables

Now we turn from color to spin. That is, we ask how to organize the spin quantum numbers of the external states in order to simplify the calculation. The answer is that the helicity basis is a very convenient one for most purposes. In high-energy collider processes, almost all fermions are ultra-relativistic, behaving as if they were massless. Massless fermions that interact through gauge interactions have a conserved helicity, which we can exploit by computing in the helicity basis. Although vector particles like photons and gluons do not have a conserved helicity, it turns out that the most helicity-violating processes one can imagine are zero at tree level (due to a hidden supersymmetry that relates boson and fermion amplitudes).

Also, the nonzero amplitudes are ordered in complexity by how much helicity violation they have; we will see that the so-called maximally helicity violating (MHV) amplitudes are the simplest, the next-to-MHV are the next simplest, and so on.

A related question is, what are the right kinematic variables for scattering amplitudes? It is traditional to use the four-momenta, $k_i^\mu$, and especially their Lorentz-invariant products, $s_{ij} = (k_i + k_j)^2$, as the basic kinematic variables. However, all the particles in the Standard Model — except the Higgs boson — have spin, and for particles with spin, there is a better choice of variables. Just as we rewrote the color factors for $SU(N_c)$ adjoint states ($f^{abc}$) in terms of those associated with the smaller fundamental representation of $SU(N_c)$ ($T^a$), we should now consider trading the Lorentz vectors $k_i^\mu$ for kinematic quantities that transform under a smaller representation of the Lorentz group.

The only available smaller representation of the Lorentz group is the spinor representation, which for massless vectors can be two-dimensional (Weyl spinors). So we trade the four-momentum $k_i^\mu$ for a pair of spinors,

$$k_i^\mu \quad \Rightarrow \quad u_+(k_i) \equiv |i^+\rangle \equiv \lambda_i^\alpha, \quad u_-(k_i) \equiv |i^-\rangle \equiv \lambda_i^{\dot{\alpha}}.$$  \hspace{1cm} (9)

Here $u_+(k_i) = \frac{1}{2}(1 + \gamma_5)u(k_i)$ is a right-handed spinor written in four-component Dirac notation, and $\lambda_i^\alpha$ is its two-component version, $\alpha = 1, 2$. Similarly, $u_-(k_i) = \frac{1}{2}(1 - \gamma_5)u(k_i)$ is a left-handed spinor in Dirac notation, and $\lambda_i^{\dot{\alpha}}$ is the two-component version, $\dot{\alpha} = 1, 2$. We also give the “ket” notation that is often used. The massless Dirac equation is satisfied by these spinors,

$$\not{k}_i u_\pm(k_i) = \not{k}_i|i^\pm\rangle = 0.$$  \hspace{1cm} (10)

There are also negative-energy solutions $v_\pm(k_i)$, but for $k_i^2 = 0$ they are not distinct from $u_\mp(k_i)$. The undotted and dotted spinor indices correspond to two different spinor representations of the Lorentz group.

We would like to build some Lorentz-invariant quantities out of the spinors, which we can do using the antisymmetric tensors $\varepsilon^{\alpha\beta}$ and $\varepsilon^{\dot{\alpha}\dot{\beta}}$. We define the spinor products,

$$\langle i j \rangle \equiv \varepsilon^{\alpha\beta}(\lambda_i)_\alpha(\lambda_j)_\beta = \bar{u}_-(k_i)u_+(k_j),$$  \hspace{1cm} (11)

$$[i j] \equiv \varepsilon^{\dot{\alpha}\dot{\beta}}(\lambda_i)^{\dot{\alpha}}(\lambda_j)^{\dot{\beta}} = \bar{u}_+(k_i)u_-(k_j).$$  \hspace{1cm} (12)
where we give both the two- and four-component versions.

Recall the form of the positive energy projector for $m = 0$:

$$u_+(k_i)\bar{u}_+(k_i) = |i^+\rangle \langle i^+| = \frac{1}{2}(1 + \gamma_5)k_i\frac{1}{2}(1 - \gamma_5).$$

(13)

In two-component notation, this relation becomes, using the explicit form of the Pauli matrices,

$$(\lambda_i)_{\alpha}(\tilde{\lambda}_i)_{\dot{\alpha}} = k^\mu_i(\sigma_\mu)_{\alpha\dot{\alpha}} = (\tilde{k}^\mu_i)_{\alpha\dot{\alpha}} = \left(\frac{k^0_i + k^z_i}{2} + ik^y_i, \frac{k^0_i + k^z_i}{2} - ik^y_i\right).$$

(14)

Note that the determinant of this $2 \times 2$ matrix vanishes, $\det(\tilde{k}^\mu_i) = k_i^2 = 0$, which is consistent with its factorization into a column vector $(\lambda_i)_{\alpha}$ times a row vector $(\lambda_i)_{\dot{\alpha}}$.

Also note that if the momentum vector $k^\mu_i$ is real, then complex conjugation is equivalent to transposing the matrix $k_i$, which via Eq. (14) corresponds to exchanging the left- and right-handed spinors, $(\lambda_i)_{\alpha} \leftrightarrow (\lambda_i)_{\dot{\alpha}}$. In other words, for real momenta, a chirality flip of all spinors (which could be induced by a parity transformation) is the same as complex conjugating the spinor products,

$$[i \ j] = \langle i \ j \rangle^*. \quad (15)$$

If we contract Eq. (14) with $(\sigma^\nu)^{\alpha\dot{\alpha}}$, we find that we can reconstruct the four-momenta $k^\mu_i$ from the spinors,

$$\langle i^+| \gamma^\mu|i^+\rangle \equiv (\tilde{\lambda}_i)_{\alpha}(\sigma^\mu)^{\alpha\dot{\alpha}}(\lambda_i)_{\alpha} = 2k^\mu_i. \quad (16)$$

Using the Fierz identity for Pauli matrices,

$$(\sigma^\mu)_{\alpha\dot{\alpha}}(\sigma_\mu)^{\beta\dot{\beta}} = 2\delta_\alpha^\beta\delta_{\dot{\alpha}}^{\dot{\beta}}, \quad (17)$$

we can similarly reconstruct the momentum invariants from the spinor products,

$$2k_i \cdot k_j = \frac{1}{2}(\tilde{\lambda}_i)_{\dot{\alpha}}(\sigma^\mu)^{\dot{\alpha}\alpha}(\tilde{\lambda}_j_{\alpha}(\lambda_i)_{\dot{\beta}}(\sigma^\mu)^{\dot{\beta}\beta}(\lambda_j)_{\beta} = (\lambda_i)_{\alpha}(\lambda_j)^{\alpha}(\tilde{\lambda}_j)_{\dot{\alpha}}(\tilde{\lambda}_i)_{\dot{\alpha}},$$

(18)

or

$$s_{ij} = \langle i \ j \rangle \langle j \ i \rangle. \quad (19)$$

For real momenta, we can combine Eqs. (15) and (19) to see that the spinor products are complex square roots of the momentum-invariants,

$$\langle i \ j \rangle = \sqrt{s_{ij}}e^{i\phi_{ij}}, \quad [i \ j] = \sqrt{s_{ij}}e^{-i\phi_{ij}},$$

(20)

where $\phi_{ij}$ is some phase. We will see later that this complex square-root property allows the spinor products to capture perfectly the singularities of amplitudes as two massless momenta become parallel (collinear). This fact is one way of understanding why helicity amplitudes can be so compact when written in terms of spinor products.

We collect here some useful spinor product identities:

- anti-symmetry : $\langle i \ j \rangle = -\langle j \ i \rangle$, $[i \ j] = -[j \ i]$, $\langle i \ i \rangle = [i \ i \rangle = 0$, (21)
- squaring : $\langle i \ j \rangle \langle j \ i \rangle = s_{ij}$, (22)
- momentum conservation : $\sum_{j=1}^{n} \langle i \ j \rangle [j \ k] = 0$, (23)
- Schouten : $\langle i \ j \rangle \langle k \ l \rangle - \langle i \ k \rangle \langle j \ l \rangle = \langle i \ l \rangle \langle k \ j \rangle$, (24)
Fig. 3: The one Feynman diagram for $e^- e^+ \to q \bar{q}$. Particles are labeled with $L$ and $R$ subscripts for left- and right-handed particles. We also give in black the numerical, all-outgoing labeling convention.

Note also that the massless Dirac equation in two-component notation follows from the antisymmetry of the spinor products:

$$ (k_i)_{\dot{\alpha} \alpha} (\lambda_i)^{\alpha} = (\tilde{\lambda}_i)_{\dot{\alpha}} \langle i \, i \rangle = 0. $$

Finally, for numerical evaluation it is useful to have explicit representations of the spinors,

$$ (\lambda_i)_{\alpha} = \left( \begin{array}{c} \sqrt{k_i^t + k_i^z} \\ k_i^x + ik_i^y \end{array} \right), \quad (\tilde{\lambda}_i)_{\dot{\alpha}} = \left( \begin{array}{c} \sqrt{k_i^t - ik_i^y} \\ k_i^x + k_i^z \end{array} \right), \quad (26) $$

which satisfy Eqs. (14) and (15).

We would like to have the same formalism describe amplitudes that are related by crossing symmetry, i.e., by moving various particles between the initial and final states. In order to keep everything on a crossing-symmetric footing, we define the momenta as if they were all outgoing, so that initial-state momenta are assigned the negative of their physical momenta. Then momentum conservation for an $n$-point process takes the crossing symmetric form,

$$ \sum_{i=1}^{n} k_i^\mu = 0. \quad (27) $$

We also label the helicity as if the particle were outgoing. For outgoing particles this label is the physical helicity, but for incoming particles it is the opposite. Because of this, whenever we look at a physical pole of an amplitude, and assign helicities to an intermediate on-shell particle, the helicity assignment will always be opposite for amplitudes appearing on two sides of a factorization pole. The same consideration will apply to particles crossing a cut, at the loop level.

3.2 A simple four-point example

Let’s illustrate spinor-helicity methods with the simplest scattering amplitude of all, the one for electron-positron annihilation into a massless fermion pair, say a pair of quarks, for which the single Feynman diagram is shown in Fig. 3. This amplitude is related by crossing symmetry to the amplitude for electron-quark scattering at the core of deep inelastic scattering, and by time reversal symmetry to the annihilation of a quark and anti-quark into a pair of leptons, i.e. the Drell-Yan reaction.

We take all the external states to be helicity eigenstates, choosing first to consider, 

$$ e_L^- (-k_1) e_R^+ (-k_2) \to q_R (k_3) \bar{q}_L (k_4). \quad (28) $$

Note that we have assigned momenta $-k_1$ and $-k_2$ to the incoming states, so that momentum conservation takes the crossing-symmetric form,

$$ k_1 + k_2 + k_3 + k_4 = 0. \quad (29) $$
In the all-outgoing helicity labeling convention, the incoming left-handed electron is labeled as if it were an outgoing right-handed positron (positive-helicity $\bar{e}$), and similarly for the incoming right-handed positron (labeled as a negative-helicity $e$). We label the amplitude with numerals $i$ standing for the momenta $k_i$, subscripts to identify the type of particle (if it is not a gluon), and superscripts to indicate the helicity. Thus the amplitude for reaction (28) is denoted by

\[ A_{4}^{\text{tree}}(1_{\bar{e}}, 2_{e}, 3_{q}, 4_{\bar{q}}) \equiv A_{4} \tag{30} \]

As discussed above, we first strip off the color factors, as well as any other coupling factors. In this case the color factor is a trivial Kronecker $\delta$ that equates the quark colors. We define the color-stripped amplitude $A_{4}$ by

\[ A_{4} = (\sqrt{2}e)^2 Q_{e}Q_{q} \delta_{i3}^{ij} A_{4}, \tag{31} \]

where $e$ is the electromagnetic coupling, obeying $e^2/(4\pi) = \alpha_{\text{QED}}$, and $Q_{e}$ and $Q_{q}$ are the electron and quark charges. The factor of $(\sqrt{2}e)^2$ arises because it is convenient to normalize the color-stripped amplitudes so that there are no $\sqrt{2}$ factors for QCD. In this normalization, the substitution $g \rightarrow \sqrt{2}e$ is required in the prefactor for each QED coupling. A corresponding $(1/\sqrt{2})^2$ goes into the Feynman rule for $A_{4}$.

The usual Feynman rules for the diagram in Fig. 3 give

\[ A_{4} = \frac{i}{2s_{12}} \gamma^\mu u_{\mu}(k_{1}) \bar{u}_{\bar{\mu}}(k_{3}) \gamma_{\mu} v_{\mu}(k_{4}) \]

\[ = \frac{i}{2s_{12}} (\sigma^\mu)_{\alpha\beta}(\lambda_{2})^\alpha(\lambda_{1})^\beta(\lambda_{3})^\gamma(\lambda_{4})^\delta, \tag{32} \]

where we have switched to two-component notation in the second line. Now we apply the Fierz identity for Pauli matrices, Eq. (17), obtaining

\[ A_{4} = \frac{i}{s_{12}}(\lambda_{2})^\alpha(\lambda_{1})^\beta(\lambda_{4})^\gamma(\lambda_{3})^\delta = \frac{i}{s_{12}} \frac{\langle 2 \ 4 \rangle [1 \ 3]}{[2 \ 4]}, \tag{33} \]

after using the definitions (11) and (12) of the spinor products $\langle i \ 4 \rangle$ and $\langle i \ 3 \rangle$.

According to Eqs. (22) and (15), the spinor products are square-roots of the momentum invariants, up to a phase. Because $s_{24} = s_{13}$ for massless four-point kinematics, we can rewrite Eq. (33) as

\[ A_{4} = \frac{i}{s_{12}} \frac{\langle 2 \ 4 \rangle [1 \ 3]}{[2 \ 4]} = e^{i\phi} \frac{s_{13}}{s_{12}} = -\frac{e^{i\phi}}{2}(1 - \cos \theta), \tag{34} \]

where $\phi$ is some phase angle, and $\theta$ is the center-of-mass scattering angle. From this formula, we can check the helicity suppression of the amplitude in the forward scattering limit, $A_{4} \rightarrow 0$ as $\theta \rightarrow 0$. The amplitude vanishes in this limit because of angular-momentum conservation: the initial angular momentum along the $e_{\bar{L}}$ direction is $(-\frac{1}{2}) - \frac{1}{2} = -1$, while the final angular momentum is $\frac{1}{2} - (-\frac{1}{2}) = +1$. At $\theta = \pi$, the spins line up and there is no suppression.

The result (33) for $A_{4}$ is in a mixed representation, containing both the “holomorphic” (right-handed) spinor product $\langle 2 \ 4 \rangle$ and the “anti-holomorphic” (left-handed) spinor product $[1 \ 3]$. However, we can multiply top and bottom by $\langle 1 \ 3 \rangle$, and use the squaring relation (22), $s_{13} = s_{24}$ and momentum conservation (23) to rewrite it as

\[ A_{4} = \frac{i}{s_{12}} \frac{\langle 2 \ 4 \rangle [1 \ 3]}{[2 \ 4]} = \frac{i}{s_{12}} \frac{\langle 2 \ 4 \rangle [1 \ 3]}{[2 \ 4]} \frac{\langle 1 \ 2 \rangle [2 \ 4]}{[1 \ 2 \ 4 \ 3]} = \frac{i}{s_{12}} \frac{\langle 2 \ 4 \rangle^2}{[1 \ 2 \ 3 \ 4]}. \tag{35} \]

The latter form only involves the spinors $\langle i \ 4 \rangle$. On the other hand, the same identities also allow us to write it in an anti-holomorphic form. In summary, we have

\[ A_{4}^{\text{tree}}(1_{\bar{e}}, 2_{e}, 3_{q}, 4_{\bar{q}}) = i \frac{\langle 2 \ 4 \rangle^2}{[1 \ 2 \ 3 \ 4]} = i \frac{[1 \ 3]^2}{[1 \ 2 \ 3 \ 4]}. \tag{36} \]
It turns out that $A_4^{\text{tree}}(1^+_e, 2^-, 3^+, 4^-)$ is the first in an infinite series of “maximally helicity violating” (MHV) amplitudes, containing these four fermions along with $(n - 4)$ additional positive-helicity gluons or photons. All of these MHV amplitudes, containing exactly two negative-helicity particles, are holomorphic. (We will compute one of them in a little while.) But $A_4^{\text{tree}}(1^+_e, 2^-, 3^+, 4^-)$ is also the first in an infinite series of $\overline{\text{MHV}}$ amplitudes, containing these four fermions along with $(n - 4)$ additional negative-helicity gluons or photons. All the MHV amplitudes are anti-holomorphic; in fact, they are the parity conjugates of the MHV amplitudes. As a four-point amplitude, eq. (36) has a dual life, belonging to both the MHV and the $\overline{\text{MHV}}$ series. The same phenomenon occurs for other classes of amplitudes, including the $n$-gluon MHV amplitudes (the Parke-Taylor amplitudes [5]) and their $\overline{\text{MHV}}$ conjugate amplitudes, which we will encounter shortly.

So far we have only computed one helicity configuration for $e^+e^- \rightarrow q\bar{q}$. There are 16 configurations in all. However, the helicity of massless fermions is conserved when they interact with gauge fields, or in the all-outgoing labeling, the positron’s helicity must be the opposite of the electron’s, and the antiquark’s helicity must be the opposite of the quark’s. So there are only $2 \times 2 = 4$ nonvanishing helicity configurations. They are all related by parity ($P$) and by charge conjugation (C) acting on one of the fermion lines. For example, C acting on the electron line exchanges labels 1 and 2, which can also be interpreted as flipping the helicities of particles 1 and 2, taking us from eq. (36) to

$$A_4^{\text{tree}}(1^-_e, 2^+_e, 3^-_q, 4^-_q) = -i \frac{(14)^2}{(12)(34)}.$$

(37)

Parity flips all helicities and conjugates all spinors, $(i j) \rightarrow [i j]$, taking us from eq. (36) to

$$A_4^{\text{tree}}(1^-_e, 2^+_e, 3^+_q, 4^-_q) = i \frac{[24]^2}{[12][34]}.$$

(38)

Combining the two operations leads to

$$A_4^{\text{tree}}(1^+_e, 2^-_e, 3^-_q, 4^+_q) = -i \frac{[14]^2}{[12][34]}.$$

(39)

Of course Eqs. (37), (38) and (39) can all be rewritten in the conjugate variables as well.

The scattering probability, or differential cross section, is proportional to the square of the amplitude. Squaring a single helicity amplitude would give the cross section for fully polarized incoming and outgoing particles. In QCD applications, we rarely have access to the spin states of the partons. Hadron beams are usually unpolarized, so the incoming quarks and gluons are as well. The outgoing quarks and gluons shower and fragment to produce jets of hadrons, wiping out almost all traces of final-state parton helicities. In other words, we need to construct the unpolarized cross section, by summing over all possible helicity configurations. (The different helicity configurations do not interfere with each other.) For our $e^+e^- \rightarrow q\bar{q}$ example, we need to sum over the four nonvanishing helicity configurations, after squaring the tree-level helicity amplitudes. The result, omitting the overall coupling and flux factors, is

$$\frac{d\sigma}{d\cos \theta} \propto \sum_{\text{hel.}} |A_4|^2 = 2 \left\{ \left| \frac{(24)^2}{(12)(34)} \right|^2 + \left| \frac{(14)^2}{(12)(34)} \right|^2 \right\}$$

$$= 2 \frac{s_{24}^2 + s_{14}^2}{s_{12}^2}$$

$$= \frac{1}{2} \left[ (1 - \cos \theta)^2 + (1 + \cos \theta)^2 \right]$$

$$= 1 + \cos^2 \theta.$$  

(40)

We used the fact that the amplitudes related by parity are equal up to a phase, in order to only exhibit two of the four nonzero helicity configurations explicitly.

40
For a simple process like $e^+ e^- \rightarrow q \bar{q}$, helicity amplitudes are overkill. It would be much faster to use the textbook method of computing the unpolarized differential cross section directly, by squaring the amplitude for generic external spinors and using Casimir’s trick of inserting the positive energy projector for the product of two spinors, summed over spin states. The problem with this method is that the computational effort scales very poorly when there a large number of external legs $n$. The number of Feynman diagrams grows like $n!$, so the number of separate interferences between diagrams in the squared amplitude goes like $(n!)^2$. That is why all modern methods for high-multiplicity scattering processes compute amplitudes, not cross sections, for some basis of external polarization states. For massless particles, this is usually the helicity basis. After computing numerical values for the helicity amplitudes at a given phase-space point, the cross section is constructed from the helicity sum.

### 3.3 Helicity formalism for massless vectors

Next we consider external massless vector particles, i.e. gluons or photons. Spinor-helicity techniques began in the early 1980s with the recognition [2] that polarization vectors for massless vector particles with definite helicity could be constructed from a pair of massless spinors, as follows:

$$
\begin{align*}
\langle e_i^+ \rangle_{\mu} = e_i^+ (k_i, q) &= \frac{\langle q^- | \gamma_\mu | i^- \rangle}{\sqrt{2} (q i)}, \quad \langle e_i^- \rangle_{\mu} = e_i^- (k_i, q) = -\frac{\langle q^+ | \gamma_\mu | i^+ \rangle}{\sqrt{2} (q i)}, \\
\langle \sigma_i^+ \rangle_{\alpha\dot{\alpha}} = \sigma_i^+ (k_i, q) &= \frac{\sqrt{2} \lambda_\alpha^i \bar{\lambda}_{\dot{\alpha}}^i}{(q i)}, \quad \langle \sigma_i^- \rangle_{\alpha\dot{\alpha}} = \sigma_i^- (k_i, q) = -\frac{\sqrt{2} \lambda_\alpha^i \bar{\lambda}_{\dot{\alpha}}^i}{(q i)},
\end{align*}
$$

where we have also given the $2 \times 2$ matrix version, from contracting with a $\sigma$ matrix and using the Fierz identity (17). Here $k_i^\mu$ is the gluon momentum and $q^\mu$ is an additional massless vector called the reference momentum, whose associated two-component left- and right-handed spinors are $\lambda_\alpha^i$ and $\bar{\lambda}_{\dot{\alpha}}^i$. Using the massless Dirac equation,

$$
\not{k_i} | i^\pm \rangle = 0 = q | q^\pm \rangle,
$$

we see that the polarization vectors (41) obey the required transversality with respect to the gluon momentum,

$$
\varepsilon_i^\pm \cdot k_i = 0.
$$

As a bonus, it also is transverse with respect to $q$: $\varepsilon_i^\pm \cdot q = 0$.

The second form (42) for the polarization vector shows that $\sigma_i^+$ produces a state with helicity $+1$, because it contains two complex conjugate spinors with momentum $k_i$ in the numerator and denominator. These two spinors pick up opposite spin-1/2 phases under an azimuthal rotation about the $k_i$ axis,

$$
\lambda_\alpha^i \rightarrow e^{i\phi/2} \bar{\lambda}_{\dot{\alpha}}^i, \quad \bar{\lambda}_{\dot{\alpha}}^i \rightarrow e^{-i\phi/2} \lambda_\alpha^i,
$$

so the ratio transforms like helicity $+1$,

$$
\sigma_i^+ \propto \frac{\lambda_\alpha^i}{\bar{\lambda}_{\dot{\alpha}}^i} \rightarrow e^{i\phi} \sigma_i^+.
$$

There is a freedom to choose different reference vectors $q_i$ for each of the external states $i$. This freedom is the residual on-shell gauge invariance, that amplitudes should be unchanged when the polarization vector is shifted by an amount proportional to the momentum. A judicious choice of the reference vectors can greatly simplify a Feynman diagram computation by causing many diagrams to vanish. However, we won’t be doing many Feynman diagram computations, just the one in the next subsection, of a five-point amplitude. In this case, there are only two diagrams, one of which we will make vanish through a choice of $q$. 


3.4 A five-point amplitude

In this subsection we compute one of the next simplest helicity amplitudes, the one for producing a gluon along with the quark-antiquark pair in $e^+e^-$ annihilation. This amplitude contributes to three-jet production in $e^+e^-$ annihilation, and to the next-to-leading order corrections to deep inelastic scattering and to Drell-Yan production, in the crossed channels.

We compute the amplitude for the helicity configuration

$$e^-_L(-k_1)e^+_R(-k_2) \to q_R(k_3)g_R(k_4)\bar{q}_L(k_5),$$

namely

$$A^{\text{tree}}_5(1^+e^-_L,2^-e^+_R,3^+q_R,4^+g_R,5^-\bar{q}_L) \equiv A_5.$$  

Again we strip off the color and charge factors, defining

$$A_5 = (\sqrt{2e})^2 g Q_e Q_q (T^{\mu_4})^5_{\mu_5} A_5,$$

where $A_5$ is constructed from the two Feynman diagrams in Fig. 4.

Recall that in the evaluation of the four-point amplitude (33), after applying the Fierz identity related to the photon propagator, the two external fermions with the same (outgoing) helicity had their spinors contracted together, generating factors of the off-shell fermion propagator and the gluon polarization vector, giving

$$A_5 = -i \frac{\langle 25 \rangle}{s_{12}} \frac{\langle 1^+ | (k_3 + k_4) \frac{g_4^\mu}{\sqrt{2s_{34}}} | 3^- \rangle}{\sqrt{2s_{34}}}.$$  

Inserting the formula (42) for the gluon polarization vector, we obtain

$$A_5 = -i \frac{\langle 25 \rangle}{s_{12}} \frac{\langle 1^+ | (k_3 + k_4) | q^+ \rangle}{s_{34}} \frac{\langle 4 \rangle}{\sqrt{2s_{45}}} + i \frac{[13]}{s_{12}} \frac{\langle 2^- | (k_4 + k_5) | q^- \rangle}{s_{45}} \frac{\langle 5 \rangle}{\langle 4 \rangle}.$$  

Now we choose the reference momentum $q = k_5$ in order to make the second graph vanish,

$$A_5 = -i \frac{\langle 25 \rangle}{s_{12}} \frac{\langle 1^+ | (k_3 + k_4) | 5^+ \rangle}{s_{34}} \frac{[43]}{\langle 5 \rangle} = -i \frac{\langle 25 \rangle}{s_{12}} \frac{[12]}{s_{34}} \frac{[21]}{[34]} \frac{[43]}{[45]} = i \frac{\langle 25 \rangle^2}{\langle 12 \rangle \langle 34 \rangle \langle 45 \rangle},$$

where we used momentum conservation (23) and a couple of other spinor-product identities to simplify the answer to its final holomorphic form,

$$A_5(1^+_e,2^-e^+_R,3^+_q,4^+g_R,5^-\bar{q}_L) = i \frac{\langle 25 \rangle^2}{\langle 12 \rangle \langle 34 \rangle \langle 45 \rangle}.$$
(As an exercise in spinor-product identities, verify Eq. (53) for other choices of $q$.)

Next we will study the behavior of $A_5$ in various kinematic limits, which will give us insight into the generic singular behavior of QCD amplitudes.

4 Soft and collinear factorization

In this section, we use the five-point amplitude (53) to verify some universal limiting behavior of QCD amplitudes. In the next section, we will use this universal behavior to derive recursion relations for general tree amplitudes.

4.1 Soft gluon limit

First consider the limit that the gluon momentum $k_4$ in Eq. (53) becomes soft, i.e. scales uniformly to zero, $k_4 \rightarrow 0$. In this limit, we can factorize the amplitude into a divergent piece that depends on the energy and angle of the emitted gluon, and a second piece which is the amplitude omitting that gluon:

$$A_5(1^+, 2^- , 3^+, 4^+, 5^-) = i \frac{\langle 25 \rangle^2}{\langle 12 \rangle \langle 34 \rangle \langle 45 \rangle} = \frac{\langle 35 \rangle}{\langle 34 \rangle \langle 45 \rangle} \times i \frac{\langle 25 \rangle^2}{\langle 12 \rangle \langle 35 \rangle} \rightarrow S(3, 4^+ , 5) \times A_4(1^+_e , 2^- , 3^+_q , 5^-).$$

The soft factor (or eikonal factor) is given more generally by,

$$S(a, s^+, b) = \frac{\langle ab \rangle}{\langle as \rangle \langle sb \rangle}, \quad S(a, s^-, b) = -\frac{[ab]}{[as][sb]},$$

where $s$ labels the soft gluon, and $a$ and $b$ label the two hard partons that are adjacent to it in the color ordering.

Although we have only inspected the soft limit of one amplitude, the more general result is,

$$A_{n}^{\text{tree}}(1, 2, \ldots , a, s^\pm , b, \ldots , n) \xrightarrow{k_s \rightarrow 0} S(a, s^\pm , b) \times A_{n-1}^{\text{tree}}(1, 2, \ldots , a, b, \ldots , n).$$

This factorization is depicted in Fig. 5. The $(n-1)$-point amplitude on the right-hand side is that obtained by just deleting the soft-gluon $s$ in the $n$-point amplitude. The soft factor is universal: it does not depend on whether $a$ and $b$ are quarks or gluons; it does not care about their helicity; and it does not even depend on the magnitude of their momenta, just their angular direction (as one can see by rescaling the spinor $\lambda_a$ in Eq. (55)). The spin independence arises because soft emission is long-wavelength, and intrinsically classical. Because of this, we can pretend that the external partons $a$ and $b$ are scalars, and
compute the soft factor simply from two Feynman diagrams, from emission off legs $a$ and $b$. We can use the scalar QED vertex in the numerator, while the (singular) soft limit of the adjacent internal propagator generates the denominator:

$$S(a, s^+, b) = -\sqrt{2} \epsilon_s^+(q) \cdot k_a + \sqrt{2} \epsilon_s^+(q) \cdot k_b = \langle a q \rangle \langle s q \rangle \langle a s \rangle - \langle b q \rangle \langle s q \rangle \langle b s \rangle = \langle a b \rangle \langle a s \rangle \langle s b \rangle,$$

(57)

using the Schouten identity (24) in the last step.

### 4.2 Collinear limits

Next consider the limit of the $e^+e^- \rightarrow qg\bar{q}$ amplitude (53) as the quark momentum $k_3 \equiv k_a$ and the gluon momentum $k_4 \equiv k_b$ become parallel, or collinear. This limit is singular because the intermediate momentum $k_P \equiv k_a + k_b$ is going on shell in the collinear limit:

$$k_P^2 = 2k_a \cdot k_b \xrightarrow{a \parallel b} 0.$$  

(58)

We also need to specify the relative longitudinal momentum fractions carried by partons $a$ and $b$,

$$k_a \approx zk_P, \quad k_b \approx (1 - z)k_P,$$

(59)

where $0 < z < 1$. This relation implies, thanks to eq. (26), that the spinors obey similar relations with square roots:

$$\lambda_a \approx \sqrt{z} \lambda_P, \quad \lambda_b \approx \sqrt{1 - z} \lambda_P,$$

(60)

$$\tilde{\lambda}_a \approx \sqrt{z} \tilde{\lambda}_P, \quad \tilde{\lambda}_b \approx \sqrt{1 - z} \tilde{\lambda}_P.$$  

(61)

Inserting Eq. (60) into Eq. (53), we find that

$$A_5(1^+, 2^-, 3^+, 4^+, 5^-) = i \frac{\langle 25 \rangle^2}{\langle 12 \rangle \langle 34 \rangle \langle 45 \rangle} \approx \frac{1}{\sqrt{1 - z}} \frac{\langle 25 \rangle^2}{\langle 12 \rangle \langle 34 \rangle} \times i \frac{\langle 25 \rangle^2}{\langle 12 \rangle \langle 34 \rangle} \times \text{Split}_-(3^+, 4^+, 5^-; z) \times A_4(1^+, 2^-, 3^+, 5^-, 4^+).$$

(62)

Here we have introduced the *splitting amplitude* $\text{Split}_{-\lambda_P}(a^{\lambda_a}, b^{\lambda_b}; z)$, which governs the general collinear factorization of tree amplitudes depicted in fig. 6,

$$A_{\text{tree}}^n(\ldots, a^{\lambda_a}, b^{\lambda_b}, \ldots) \xrightarrow{a \parallel b} \sum_{\lambda_P = \pm} \text{Split}_{-\lambda_P}(a^{\lambda_a}, b^{\lambda_b}; z) A_{\text{tree}}^{n-1}(\ldots, P^{\lambda_P}, \ldots).$$

(63)
In contrast to the soft factor, the splitting amplitude depends on whether $a$ and $b$ are quarks or gluons, and on their helicity. It also includes a sum over the helicity $\lambda_P$ of the intermediate parton $P$. (Note that the labeling of $\lambda_P$ is reversed between the $(n - 1)$-point tree amplitude and the splitting amplitude, because we apply the all-outgoing helicity convention to the splitting amplitude as well.) The $(n - 1)$-point tree amplitude on the right-hand side of Eq. (63) is found by merging the two partons, according to the possible splittings in QCD: $g \to gg$, $g \to q\bar{q}$, $\bar{q} \to \bar{q}g$, and (in this case) $q \to gg$. For the splitting amplitude $\text{Split}_- (a^{+}_{+}, b^{+}_{-}; z)$ entering Eq. (62), quark helicity conservation implies that only one of the two intermediate helicities survives. For intermediate gluons, both signs of $\lambda_P$ can appear in general. As in the case of the soft limit, the four-point amplitude $A_4$ is found by relabeling Eq. (36).

One can also extract from Eq. (53) the splitting amplitude for the case that the (anti)quark and gluon have the opposite helicity, by taking the collinear limit $4 \parallel 5$. The two results can be summarized as:

$$\text{Split}_- (q^+, g^-) = \frac{1}{\sqrt{1 - z\langle qg \rangle}}, \quad \text{Split}_- (q^+, g^+) = -\frac{z}{\sqrt{1 - z\langle q\bar{g} \rangle}}, \quad \text{Split}_- (g^+, q^-) = \frac{1}{\sqrt{z\langle g\bar{q} \rangle}}, \quad \text{Split}_- (g^+, q^+) = -\frac{1 - z}{\sqrt{z\langle g\bar{q} \rangle}},$$

(64)

(65)

where the other cases (including some not shown, with opposite quark helicity) are related by parity or charge conjugation.

Collinear singularities in the initial state give rise to the DGLAP evolution equations for parton distributions. In fact, the splitting amplitudes are essentially the square root of the (polarized) Altarelli-Parisi splitting probabilities which are the kernels of the DGLAP equations. That is, the $z$ dependence of the splitting amplitudes, after squaring and summing over the helicities $\lambda_a$, $\lambda_b$, and $\lambda_P$, reproduces the splitting probabilities. For example, one can reconstruct the correct $z$-dependence of the $q \to qg$ splitting probabilities $P_{qq}(z)$ using Eq. (64), squaring and summing over the gluon helicity:

$$P_{qq}(z) \propto \left( \frac{1}{\sqrt{1 - z}} \right)^2 + \left( \frac{z}{\sqrt{1 - z}} \right)^2 = 1 + z^2, \quad \frac{1}{1 - z},$$

(66)

while $P_{gq}(z)$ is given by exchanging $z \leftrightarrow 1 - z$. Equation (66) omits the $\delta(1 - z)$ term from virtual gluon emission, but its coefficient can be inferred from quark number conservation.

### 4.3 The Parke-Taylor amplitudes

In the all-outgoing helicity convention, one can show that the pure-gluon amplitudes for which all the gluon helicities are the same, or at most one is different from the rest, vanish for any $n \geq 4$:

$$A^\text{tree}_n(1^+, 2^+, \ldots, n^+) = 0.$$  

(67)

(Cyclic symmetry allows us to move a single negative-helicity gluon to leg 1.) This result can be proven directly by noticing that the tree amplitude contains $n$ different polarization vectors, contracted together with at most $n - 2$ momenta (because there are at most $n - 2$ cubic vertices in any Feynman graph, each of which is linear in the momentum). Therefore every term in every tree amplitude contains at least one polarization vector contraction of the form $\varepsilon_i \cdot \varepsilon_j$. Inspecting the form of the polarization vectors in Eq. (41), we see that like-helicity contractions, $\varepsilon_i^+ (q_i) \cdot \varepsilon_j^+ (q_j)$, vanish if $q_i = q_j$, while opposite helicity contractions, $\varepsilon_i^- (q_i) \cdot \varepsilon_j^+ (q_j)$, vanish if $q_i = k_j$ or $q_j = k_i$. To show that $A^\text{tree}_n(1^+, 2^+, \ldots, n^+) \neq 0$, we can just choose all reference momenta to be the same, $q_i = q$. To show that $A^\text{tree}_n(1^-, 2^+, \ldots, n^+) \neq 0$, we can choose $q_i = k_i$ for $i > 1$ and $q_1 = k_2$, for example. It is also possible to prove Eq. (67) using the fact that tree-level $n$-gluon amplitudes are the same in QCD as in a supersymmetric theory [4], and so they obey Ward identities for supersymmetric scattering amplitudes [3].
The remarkable simplicity of gauge-theory scattering amplitudes is encapsulated by the Parke-Taylor \([5]\) amplitudes for the MHV \(n\)-gluon amplitudes, in which exactly two gluons, \(j\) and \(l\), have opposite helicity from the rest:

\[
A_{jl}^{\text{MHV}} \equiv A_n^{\text{tree}}(1^+, \ldots, j^-, \ldots, l^-, \ldots, n^+) = i \frac{\langle j \ l \rangle^4}{\langle 1 \ 2 \rangle \cdots \langle n \ 1 \rangle}. \tag{68}
\]

One of the reasons these amplitudes are so simple is that they have no multi-particle poles — no factors of \(1/(k_m + k_{m+1} + \cdots + k_p)^2 \equiv 1/P^2\) for \(p > m + 1\). Why is that? A multi-particle pole would correspond to factorizing the scattering process into two subprocesses, each with at least four gluons,

\[
A_n^{\text{tree}}(\ldots) \xrightarrow{P^2 \to 0} A_{n-k+1}(\ldots, P^{\lambda_P}, \ldots) \frac{i}{P^2} A_{k+1}(\ldots, (-P)^{-\lambda_P}, \ldots), \quad 3 \leq k \leq n-3, \tag{69}
\]

In the MHV case, there are two negative-helicity gluons among the arguments “…” of the two tree amplitudes on the right-hand side of eq. (69), plus one more for either \(P\) or \((-P)\) (but not both). That’s three negative-helicity gluons to be distributed among two tree amplitudes. However, Eq. (67) says that both trees need at least two negative helicities to be nonvanishing, for a minimum of four required. Hence the multiparticle poles must all vanish, due to insufficiently many negative helicities. As we’ll see in Section 6, similar arguments control the structure of loop amplitudes as well.

We have found that the MHV amplitudes have no multi-particle factorization poles, consistent with Eq. (68). Their principal singularities are the soft and collinear limits. It’s easy to check that the soft limit (56) is satisfied by the MHV amplitudes in Eq. (68). It’s also simple to verify that the collinear behavior (63) is obeyed, and to extract the \(g \to gg\) splitting amplitudes,

\[
\text{Split}_{-}(a^+, b^+) = \frac{1}{\sqrt{z(1-z)} \langle ab \rangle}, \quad \text{Split}_{+}(a^-, b^+) = \frac{z^2}{\sqrt{z(1-z)} \langle ab \rangle},
\]

\[
\text{Split}_{+}(a^+, b^-) = \frac{z^2}{\sqrt{z(1-z)} \langle ab \rangle}, \quad \text{Split}_{+}(a^+, b^+) = 0, \tag{70}
\]

plus their parity conjugates. The last relation in Eq. (70) must hold for consistency, because otherwise the collinear limit of an MHV amplitude (which has no multi-particle poles) could generate a next-to-MHV amplitude with three negative helicities (which generically does have such poles). It’s a useful exercise to reconstruct the unpolarized \(g \to gg\) splitting probabilities \(P_{gg}(z)\) from Eq. (70) by squaring and summing over all helicity configurations.

A closely related series of MHV amplitudes to the pure-glue ones are those with a single external \(q\bar{q}\) pair and \((n - 2)\) gluons. In this case helicity conservation along the fermion line forces either the quark or antiquark to have negative helicity. Using charge conjugation, we can pick it to be the antiquark. Referring to the color decomposition (6), the partial amplitudes for which all gluons have the same helicity vanish identically,

\[
A_n^{\text{tree}}(1_\bar{q}, 2_\bar{q}^+, 3^+, 4^+, \ldots, n^+) = 0, \tag{71}
\]

while the MHV ones with exactly one negative-helicity gluon (leg \(i\)) take the simple form,

\[
A_n^{\text{tree}}(1_\bar{q}, 2_\bar{q}^+, 3^+, \ldots, i^-, \ldots, n^+) = i \frac{\langle 1 \ i \rangle^3 \langle 2 \ i \rangle}{\langle 1 \ 2 \rangle \cdots \langle n \ 1 \rangle}. \tag{72}
\]

It’s easy to see that the absence of multi-particle poles in eq. (68), whether for intermediate gluons or quarks, again follows from the vanishing relations (67) and (71), and simple counting of negative helicities. However, the relation between the pure-glue MHV amplitudes \(A_{1i}^{\text{tree, MHV}}\) in Eq. (68) and the quark-glue ones (72) is much closer than that, as they differ only by a factor of \(\langle 2 \ i \rangle / \langle 1 \ i \rangle\). These relations follow from supersymmetry Ward identities \([3, 4, 15, 16]\).
Fig. 7: In gauge theory, an angular-momentum mismatch lessens the singular behavior from $1/p^2$ to $1/\sqrt{p^2}$, and introduces an azimuthally-dependent phase, both of which are captured by the spinor products.

### 4.4 Spinor magic

All of the splitting amplitudes contain denominator factors of either $\langle a \ b \rangle$ or its parity conjugate $[a \ b]$. From Eq. (20), we see that the collinear singularity is proportional to the square root of the momentum invariant that is vanishing, times a phase. This phase varies as the two collinear partons are rotated in the azimuthal direction about their common axis. Both the square root and the phase behavior follow from angular momentum conservation in the collinear limit. Figure 7 illustrates the difference between scalar $\phi^3$ theory and gauge theory. In scalar $\phi^3$ theory, no spin angular momentum is carried by either the external scalars or the intermediate one. Thus there is no violation of angular-momentum conservation along the collinear axis. Related to this, the three-vertex shown carries no momentum dependence, and the collinear pole is determined solely by the scalar propagator to be $\sim 1/s_{ab}$ in the limit that legs $a$ and $b$ become parallel.

In contrast, in every collinear limit in massless gauge theory, angular momentum conservation is violated by at least one unit. In the pure-glue case shown in Fig. 7, the intermediate physical gluon must be transverse and have helicity $\pm 1$, but this value is never equal to the sum of the two external helicities: $\pm 1 \pm 1 = \pm 2$ or 0. The helicity mismatch forces the presence of orbital angular momentum, which comes from the momentum dependence in the gauge-theory three-vertex. It suppresses the amplitude in the collinear limit, from $1/s_{ab}$ to $1/\sqrt{s_{ab}}$, similarly to the vanishing of $A_1$ in Eq. (34) in the limit $\theta \to 0$. The helicity mismatch also generates the azimuthally-dependent phase. The sign of the mismatch, by $\pm 1$ unit, is correlated with whether the splitting amplitude contains $1/\langle a \ b \rangle$ or $1/[a \ b]$, since these spinor products acquire opposite phases under an azimuthal rotation.

In summary, the spinor products are the perfect variables for capturing the collinear behavior of massless gauge theory amplitudes, simply due to angular-momentum considerations. Because collinear singularities dictate many of the denominator factors that should appear in the analytic representations of amplitudes, we can now understand more physically why the spinor product representation can lead to such compact analytic results.

### 4.5 Complex momenta, spinor products and three-point kinematics

There is another reason the spinor products are essential for modern amplitude methods, and that is to make sense out of massless three-point scattering. If we use only momentum invariants, then the three-point kinematics, defined by

$$k_1^\mu + k_2^\mu + k_3^\mu = 0, \quad k_1^2 = k_2^2 = k_3^2 = 0,$$

(73)

is pathological. For example, $s_{12} = (k_1 + k_2)^2 = k_3^2 = 0$, and similarly every momentum invariant $s_{ij}$ vanishes. If the momenta are real, then eq. (20) implies that all the spinor products vanish as well,
\( (i,j) = [i,j] = 0 \). It is easy to see that for real momenta the only solutions to Eq. (73) consist of strictly parallel four-vectors, which is another way of seeing why all dot products and spinor products must vanish.

However, if the momenta are complex, there is a loophole: The conjugation relation (15), \([i,j] = (i,j)^*\), does not hold, although the relation (19), \(s_{ij} = \langle i,j \rangle [j,i]\), is still true. Therefore we can have some of the spinor products be nonzero, even though all the momentum invariants vanish, \(s_{ij} = 0\). There are two chirally conjugate solutions:

1. \( \tilde{\lambda}_1 \propto \tilde{\lambda}_2 \propto \tilde{\lambda}_3 \Rightarrow \text{all } [i,j] = 0 \text{ while all } \langle i,j \rangle \neq 0 \).
2. \( \lambda_1 \propto \lambda_2 \propto \lambda_3 \Rightarrow \text{all } \langle i,j \rangle = 0 \text{ while all } [i,j] \neq 0 \).

The proportionality of the two-component spinors causes the corresponding spinor products to vanish.

There are no continuous variables associated with the three-point process, so one should think of the kinematical region as consisting of just two points, which are related to each other by parity.

For the first choice of kinematics, MHV three-point amplitudes such as

\[
A^\text{tree}_3(1^-,2^-,3^+) = i \frac{(12)^4}{(12)(23)(31)}
\]

(74)

make sense and are nonvanishing. MHV three-point amplitudes such as

\[
A^\text{tree}_3(1^+,2^+,3^-) = -i \frac{[12]^4}{[12][23][31]}
\]

(75)

are nonvanishing for the second type of kinematics. When the MHV three-point amplitudes are nonvanishing, the MHV ones vanish, and vice versa.

It’s important to note that the splitting amplitudes defined in section 4.2 correspond to approximate three-point kinematics with real momenta, whereas the three-point amplitudes (74) and (75) correspond to exact three-point kinematics with complex momenta. They are similar notions, but not exactly the same thing.

5 The BCFW recursion relation for tree amplitudes

5.1 General formula

The idea behind the derivation of the BCFW recursion relation [25] is that tree-level amplitudes are plastic, or continuously deformable, analytic functions of the scattering momenta. Therefore, it should be possible to reconstruct amplitudes for generic scattering kinematics from their behavior in singular limiting kinematics. In these singular regions, amplitudes split, or factorize, into two causally disconnected amplitudes with fewer legs, connected by a single intermediate state, which can propagate an arbitrary distance because it is on its mass shell.

Multi-leg amplitudes depend on many variables, and multi-variable complex analysis can be tricky. However, BCFW considered a family of on-shell tree amplitudes, \(A_n(z)\), depending on a single complex parameter \(z\) which shifts some of the momenta. (We drop the “tree” superscript here for convenience.) This family explores enough of the singular kinematical configurations to allow recursion relations to be derived for the original amplitude at \(z = 0\), \(A_n = A_n(0)\). There have since been many generalizations of this approach, leading to different types of recursion relations. The BCFW momentum shift only affects two of the momenta, say legs \(n\) and 1. The shift can be defined using the spinor variables as,

\[
\tilde{\lambda}_n \rightarrow \tilde{\lambda}_n = \tilde{\lambda}_n - z\tilde{\lambda}_1, \quad \lambda_n \rightarrow \lambda_n,
\]

\[
\lambda_1 \rightarrow \lambda_1 = \lambda_1 + z\lambda_n, \quad \tilde{\lambda}_1 \rightarrow \tilde{\lambda}_1,
\]

(76)
where hatted variables indicate variables after the shift. This particular shift is called the \([n, 1]⟩\) shift, because it only affects the spinor products involving the left-handed spinor \(\tilde{\lambda}_n\) and the right-handed spinor \(\lambda_1\).

The shift (76) can also be expressed in terms of momentum variables,

\[
\begin{align*}
\hat{k}_1(z) &= (\lambda_1 + z\lambda_n) \tilde{\lambda}_1 = \lambda_1 \tilde{\lambda}_1 + z\lambda_n \tilde{\lambda}_1, \\
\hat{k}_n(z) &= \lambda_n (\tilde{\lambda}_n - z\tilde{\lambda}_1) = \lambda_n \tilde{\lambda}_n - z\lambda_n \tilde{\lambda}_1,
\end{align*}
\]

which makes clear that momentum conservation holds for any value of \(z\), because

\[
\hat{k}_1^\mu(z) + \hat{k}_n^\mu(z) = k_1^\mu + k_n^\mu.
\]

Also, since both \(\hat{k}_1(z)\) and \(\hat{k}_n(z)\) in Eq. (77) can be factorized as \(2 \times 2\) matrices into row vectors times column vectors, their determinants vanish. Then, according to the discussion around Eq. (14), they remain on shell,

\[
\hat{k}_1^2(z) = \hat{k}_n^2(z) = 0.
\]

We can give a physical picture of the direction of the momentum shift by first writing \(A_n(z)\) in terms of momentum variables, their determinants vanish. Then, according to the discussion around Eq. (14), they remain on shell.

The function \(A_n(z)\) depends meromorphically on \(z\). If it behaves well enough at infinity, then we can use Cauchy’s theorem to relate its behavior at \(z = 0\) (the original amplitude) to its residues at finite values of \(z\) (the factorization singularities). If \(A_n(z) → 0\) as \(z → ∞\), then we have,

\[
0 = \frac{1}{2\pi i} \oint_C dz \frac{A_n(z)}{z} = A_n(0) + \sum_k \text{Res} \left[ \frac{A_n(z)}{z} \right]_{z = z_k},
\]

where \(C\) is the circle at infinity, and \(z_k\) are the locations of the factorization singularities in the \(z\) plane. (See fig. 8.) These poles occur when the amplitude factorizes into a subprocess with momenta \((k_1, k_2, \ldots, k_k, -K_{1,k})\), where \(K_{1,k}(z_k) = k_1(z_k) + k_2 + \cdots + k_k\) must be on shell. This information lets us write a simple equation for \(z_k\),

\[
0 = K_{1,k}^2(z_k) = (k_1(z_k) + k_2 + \cdots + k_k)^2 = (z_k\lambda_n \tilde{\lambda}_1 + K_{1,k})^2 = z_k \langle n^- | K_{1,k} | 1^- \rangle + K_{1,k}^2,
\]

where \(K_{1,k} = k_1 + k_2 + \cdots + k_k\). The solution to eq. (81) is

\[
z_k = -\frac{K_{1,k}^2}{\langle n^- | K_{1,k} | 1^- \rangle}.
\]

We also have to compute the residue of \(A(z)/z\) at \(z = z_k\). To do that we use Eq. (69), which also holds for three-point factorizations in complex kinematics. The singular factor in the denominator that produces the residue is

\[
zP^2(z) = zK_{1,k}^2(z) ≈ z_k \langle n^- | K_{1,k} | 1^- \rangle (z - z_k) ≈ -K_{1,k}^2 (z - z_k).
\]

Thus after taking the residue it contributes a factor of the corresponding scalar propagator, \(i/K_{1,k}^2\), evaluated for the original unshifted kinematics where it is nonsingular.

Solving Eq. (80) for \(A_n(0)\) then gives the final BCFW formula [25],

\[
A_n(1, 2, \ldots, n) = \sum_{h = ±} \sum_{k = 2}^{n-2} A_{k+1}(1, 2, \ldots, k, -K_{1,k}^{-h}) \frac{i}{K_{1,k}^2} A_{n-k+1}(K_{1,k}^{-h}, k + 1, k + 2, \ldots, n - 1, n),
\]

(84)
Fig. 8: Illustration of how Cauchy’s theorem leads to the BCFW recursion relation. The magenta dot represents the residue at the origin; the blue dots the residues at \( z_k \). In the recursion relation, the red lines carry complex, shifted momenta.

Fig. 9: Large \( z \) dependence of a generic Feynman diagram, for the \([n^-, 1^+]\) momentum shift. Only the red gluons carry the large momentum proportional to \( z^\mu \). The red

where the hat in the \( k \)th term indicates that the shifted momentum is to be evaluated for \( z = z_k \), and \( h = \pm \) labels the sign of the helicity of the intermediate state carrying (complex) momentum \( \hat{K}_{1,k} \). The sum is over the \( n - 3 \) ordered partitions of the \( n \) momenta into two sets, with at least a three-point amplitude on the left \( (k \geq 2) \) and also on the right \( (k \leq n - 2) \). The recursion relation is depicted in Fig. 8.

In order to finish the proof of Eq. (84), we need to show that \( A_n(z) \) vanishes as \( z \to \infty \). We will do so for the case that leg \( n \) has negative helicity and leg 1 has positive helicity, the so-called \([-, +]\) case. This case can be demonstrated using Feynman diagrams [25]. The cases \([+, +]\) and \([-, -]\) also vanish at infinity, but the proof is slightly more involved. The case \([+, -]\) diverges at infinity, so it should not be used as the basis for a recursion relation. Consider the large \( z \) behavior of the generic Feynman diagram shown in Fig. 9. Only the red gluons carry the large momentum proportional to \( z v^\mu \). The red
propagators contribute factors of the form
\[
\frac{1}{K_{1,k}^2(z)} = \frac{1}{K_{1,k}^2 + z \langle n^- | \hat{K}_{1,k} | 1^- \rangle} \sim \frac{1}{z}, \quad \text{as } z \to \infty. \quad (85)
\]

Yang-Mills vertices are (at worst) linear in the momentum, so they contribute a factor of \(z\) per vertex. There is one more vertex than propagator, so the amplitude scales like \(z + 1\) before we take into account the external polarization vectors. For the \([-,-]\) case, they scale like
\[
\hat{g}_n^-(q) \propto \frac{\lambda_n \lambda_q}{[n \, q]} \propto \frac{1}{z}, \quad \hat{g}_1^+(q) \propto \frac{\lambda_1 \lambda_q}{(1 \, q)} \propto \frac{1}{z}. \quad (86)
\]
The two factors of \(1/z\), combined with the factor of \(z\) from the internal part of the diagram, mean that every Feynman diagram falls off like \(1/z\), so \(A_n(\infty) = 0\) for the \([-,-]\) shift.

It is easy to see that flipping either helicity in Eq. (86) results in a polarization vector that scales like \(z\) instead of \(1/z\), invalidating the argument based on Feynman diagrams. However, it is possible to show [26] using the background field method that the \([+,+\rangle\) and \([-,-\rangle\) cases are actually just as well behaved as the \([-,-\rangle\) case, also falling off like \(1/z\). In contrast, the \([+,-\rangle\) case does diverge like \(z^3\), as suggested by the above diagrammatic argument.

### 5.2 Application to MHV

Next we apply the BCFW recursion relation to prove the form of the Parke-Taylor amplitudes (68), inductively putting one of the two negative helicities in the \(n\)th position,
\[
A_{j,n}^{\text{MHV}} = A_n^{\text{tree}}(1^+, 2^+, \ldots, j^-, \ldots, n^-) = i \frac{\langle j \, n \rangle^4}{\langle 1 2 \ldots n \rangle}. \quad (87)
\]
First we note that the middle terms in the sum over \(k\) in Eq. (84), with \(3 \leq k \leq n-3\) all vanish. That’s because they correspond to the multi-particle pole factorizations considered in Eq. (69), with at least a four-point amplitude on each side of the factorization pole, and vanish according to the discussion below Eq. (69), by counting negative helicities.

The case \(k = n-2\) also vanishes. If \(j = n-1\), then it vanishes because \(A_{k+1}\) can have at most one negative helicity. If \(j < n-1\), then we must have \(h = +\) so that \(A_{k+1}\) is non-vanishing, and then the three-point amplitude \(A_{n-k+1}\) is of type \((+,+,,-\)). This amplitude, given in Eq. (75), can be nonvanishing when the three right-handed spinors \(\lambda_i (i = K, n-1, n)\) are proportional (the second choice of three-point kinematics). However, we have shifted the left-handed spinor \(\tilde{\lambda}_n\), not the right-handed one, and it is easy to check that the three-point configuration we arrived at is the one for which three left-handed spinors \(\tilde{\lambda}_i\) are proportional. For this choice \(A_{n-k+1}\) vanishes.

The only nonvanishing contribution is from \(k = 2\). We assume \(j > 2\) for simplicity. Since we have shifted \(\lambda_1\), the three right-handed spinors \(\lambda_i (i = K, 1, 2)\) must be proportional, which allows the following three-point amplitude to be non-vanishing:
\[
A_3(1^+, 2^+, -\hat{K}^-) = -i \frac{[1 \, 2]^4}{[1 \, 2][2 (-\hat{K})][(-\hat{K}) \, 1]} = + i \frac{[1 \, 2]^3}{[2 \hat{K}][\hat{K} \, 1]}, \quad (88)
\]
where \(\hat{K} = \hat{K}_{1,2}\). We removed the hats on 1 in the second step, since \(\tilde{\lambda}_1\) is not shifted. There are also two factors of \(i\) from reversing the sign of \(\hat{K}\) in the spinor products.

The other amplitude appearing in the \(k = 2\) term in Eq. (84) is evaluated using induction on \(n\) and Eq. (87):
\[
A_{n-1}(\hat{K}^+, 3^+, \ldots, j^-, \ldots, n^-) = i \frac{\langle j \, \hat{n} \rangle^4}{\langle \hat{K} \, 3 \rangle \langle 3 \, 4 \rangle \cdots \langle n-1, \hat{n} \rangle \langle \hat{n} \, \hat{K} \rangle}.
\]
where we can again remove the hats on \( n \) because \( \lambda_n \) is unshifted.

Combining the three factors in the \( k = 2 \) term in the BCFW formula (Eq. (84)) gives

\[
A_{jn}^{\text{MHV}} = -i \frac{(j n)^4}{\langle K 3 \rangle \langle 3 4 \rangle \cdots \langle n - 1, n \rangle \langle n K \rangle} \frac{1 [12]^3}{s_{12} [2K][K1]}.
\]  

One can combine the \( \hat{K} \)-containing factors into \( \langle n \hat{K} \rangle \hat{K} 2 \) and \( \langle 3 \hat{K} \rangle \hat{K} 1 \). At this point, we would normally need the value of \( z_k \) to proceed. From Eq. (82), it is

\[
z_2 = -\frac{s_{12}}{\langle n^\perp \rangle (1 + 2) \{1 - \}} = \frac{1}{\langle n 2 \rangle^2 \{1 - \}} = \frac{1}{\langle n 2 \rangle^2}.
\]

However, the evaluation of the \( \hat{K} \)-containing strings in this case, where

\[
\hat{K} = \hat{K}_{1,2}(z_2) = \hat{k}_1 + \hat{k}_2 + z_2 \lambda_n \hat{\lambda}_1,
\]
does not actually require the value of \( z_2 \):

\[
\begin{aligned}
\langle n \hat{K} \rangle \hat{K} 2 &= \langle n^\perp \rangle (1 + 2) \{2 - \} + z_2 \langle n n \rangle [12] = \langle n 1 \rangle \{1 - \}, \\
\langle 3 \hat{K} \rangle \hat{K} 1 &= \langle 3^\perp \rangle (1 + 2) \{1 - \} + z_2 \langle 3 n \rangle [11] = \langle 3 2 \rangle \{1 - \}.
\end{aligned}
\]

Inserting these results into Eq. (90) gives

\[
A_{jn}^{\text{MHV}} = -i \frac{(j n)^4 [12]^3}{\langle 12 \rangle \langle 21 \rangle \langle 23 \rangle \langle n 1 \rangle \{1 - \}} \frac{1}{\langle n 1 \rangle \langle 3 4 \rangle \cdots \langle n - 1, n \rangle \langle n 1 \rangle},
\]

completing the induction and proving the Parke-Taylor formula.

### 5.3 An NMHV application

Now we know all the MHV pure-gluon tree amplitudes with exactly two negative helicities, and by parity, all the MHV amplitudes with exactly two positive helicities. The first gluonic amplitude which is not zero or one of these is encountered for six gluons, with three negative and three positive helicities, the next-to-MHV case. In fact, there are three inequivalent cases (up to cyclic permutations and reflection symmetries):

\[
A_6(1^+, 2^+, 3^+, 4^-, 5^-, 6^-), \quad A_6(1^+, 2^+, 3^-, 4^+, 5^-, 6^-), \quad A_6(1^+, 2^-, 3^+, 4^-, 5^+, 6^-).
\]

One can use a simple group theory relation known as the \( U(1) \) decoupling identity to rewrite the third configuration in terms of the first two [15, 16].

Here we will give a final illustration of the BCFW recursion relation by computing the first of the amplitudes in Eq. (95). (The other two are almost as simple to compute.) We again use the \( |n^-, 1^+\rangle \) shift, for \( n = 6 \). The \( k = 3 \) term vanishes in this case because \( A_{k+1} = A_4(1^+, 2^+, 3^+, -\hat{K}_{1,3}) = 0 \). The \( k = 2 \) and \( k = 4 \) terms are related by the following parity symmetry:

\[
1) \leftrightarrow 6], \quad 2) \leftrightarrow 5], \quad 3) \leftrightarrow 4], \quad 4) \leftrightarrow 3], \quad 5) \leftrightarrow 2], \quad 6) \leftrightarrow 1].
\]  

52
For the \( k = 2 \) term, using \( z_2 \) from Eq. (91), we have the kinematical identities (where again \( \hat{K} = \hat{K}_{1,2} \)),
\[
\hat{K} = k_1 + k_2 - \frac{\{1 2\}}{\{6 2\}} [6] [1], \tag{97}
\]
\[
\hat{\hat{1}} = [1], \tag{98}
\]
\[
[6] = [6] + \frac{\{1 2\}}{\{6 2\}} [1]. \tag{99}
\]

The \( k = 2 \) BCFW diagram is
\[
T_2 = A_3(\hat{1}^+, 2^+, -\hat{K}_{1,2}) \frac{i}{s_{12}} A_5(\hat{K}_{1,2}^+, 3^+, 4^-, 5^-, 6^-)
= \frac{i}{s_{12}} \left[ \frac{[2 \hat{K}][\hat{K} 1][3 4] [4 5][5 6] [6 \hat{K}]}{[\hat{K} 3]^3} \right]
= \frac{i}{s_{12}} \left[ \frac{[2 \hat{K}][\hat{K} 6)](\hat{K} 1)[3 4] [4 5][5 6] ([6 \hat{K}][\hat{K} 6]) \right]. \tag{100}
\]

Using Eqs. (97) and (99), we can derive the identities,
\[
\langle 6 \hat{K} 6 \rangle \langle K a \rangle = \langle 6^- | (1 + 2) | a^- \rangle,
\]
\[
\langle 6 1 | 3 4 | 5 6 \rangle = \frac{\{1 2\}[5 1]}{\{6 2\}} = \frac{\langle 2^- | (6 + 1) | 5^- \rangle}{\langle 6 2\rangle},
\]
\[
\langle 6 \hat{K} \hat{K} 6 \rangle = \langle 6^+ | (1 + 2) | 6^+ \rangle + s_{12} = s_{612} \tag{101}
\]
where \( s_{612} = (k_6 + k_1 + k_2)^2 \). Inserting these identities into Eq. (100) for \( T_2 \), we have
\[
T_2 = \frac{\langle 6^- | (1 + 2) | 3^- \rangle^3}{\langle 6 1 | 3 4 | 5 6 \rangle s_{612} \langle 2^- | (6 + 1) | 5^- \rangle}. \tag{102}
\]

We can use the parity symmetry (96) to obtain the \( k = 4 \) term. The final result for the six-point NMHV amplitude is,
\[
A_6(1^+, 2^+, 3^+, 4^-, 5^-, 6^-) = \frac{i}{s_{612} \langle 6 1 | 3 4 | 5 6 \rangle s_{612} \langle 2^- | (6 + 1) | 5^- \rangle}
\]
\[
+ \frac{i}{\langle 2 3 | 3 4 | 5 6 | 6 1 \rangle s_{561} \langle 2^- | (6 + 1) | 5^- \rangle}. \tag{103}
\]

It’s worth comparing the analytic form of this result to that found in the 1980’s [22],
\[
A_6(1^+, 2^+, 3^+, 4^-, 5^-, 6^-) = i \frac{(1 2) [4 5] \langle 6^- | (1 + 2) | 3^- \rangle^2}{s_{612} s_{123} s_{123} s_{123} s_{123} s_{123}}
\]
\[
+ \frac{i}{s_{23} s_{23} s_{23} s_{23} s_{23} s_{23}} \frac{(2 3) [5 6] \langle 4^- | (2 + 3) | 1^- \rangle^2}{s_{561} s_{561} s_{561} s_{561} s_{561} s_{561}}
\]
\[
+ \frac{i}{s_{123} [1 2] [2 3] [4 5] [5 6] \langle 6^- | (1 + 2) | 3^- \rangle \langle 4^- | (2 + 3) | 1^- \rangle}{s_{123} s_{123} s_{123} s_{123} s_{123} s_{123}}. \tag{104}
\]

Although the new form has only one fewer term, it represents the physical singularities in a cleaner fashion. For example, in the collinear limit \( 3 \parallel 4 \), Eq. (103) makes manifest the \( 1/\langle 3 4 \rangle \) and \( 1/\langle 3 4 \rangle \) singularities, which correspond to the two different intermediate gluon helicities that contribute in this collinear channel, as the six-point NMHV amplitude factorizes on both the MHV and MHV five-point amplitudes, \( A_5(1^+, 2^+, 3^+, 4^-, 5^-, 6^-) \). On the other hand, each term of Eq. (104) behaves like the product
of these two singularities, since \(1/s_{3,4} = -1/(⟨3,4⟩[3,4])\). Hence there are large cancellations between the three terms in this channel. Such cancellations can lead to large losses in numerical precision due to round-off errors, especially in NLO calculations which typically evaluate tree amplitudes repeatedly close to the collinear poles.

On the other hand, eq. (103) contains a spurious singularity that Eq. (104) does not, as \((2^-|(6 + 1)|5^-) \rightarrow 0\). This can happen, for example, whenever \(k_6 + k_1\) is a linear combination of \(k_2\) and \(k_5\). (In the collision \(2 + 5 \rightarrow 6 + 1 + 3 + 4\), such a configuration is reached if the vectors \(k_6 + k_1\) and \(k_3 + k_4\) have no component transverse to the beam axis defined by \(k_2\) and \(k_5\); that is, if \(k_6 + k_1\) is a linear combination of \(k_2\) and \(k_5\).) It’s called a spurious singularity because the amplitude should evaluate to a finite number there, but individual terms blow up. However, these singularities tend to have milder consequences, as long as they appear only to the first power, as they do here. That’s because the amplitude is not particularly large in this region, so in the evaluation of an integral containing it by importance-sampling, it is rare to come close enough to the surface where \((2^-|(6 + 1)|5^-)\) vanishes that round-off error is a problem. Different choices of BCFW shifts lead to different spurious singularities, so one can always check the value of \((2^-|(6 + 1)|5^-)\) and use a different shift if it is too small.

In general, the BCFW recursion relation leads to very compact analytic representations for tree amplitudes. The relative simplicity with respect to previous analytic approaches becomes much more striking for seven or more external legs. A closely related set of recursion relations for \(\mathcal{N} = 4\) super-Yang-Mills theory [27] have been solved in closed form for an arbitrary number of external legs [28]. These solutions can also be used to compute tree amplitudes, in particular, off-shell recursion relations based on the Dyson-Schwinger equations, such as the Berends-Giele recursion relations [6]. At very high multiplicities, these can be numerically even more efficient than the BCFW recursion relations. Nevertheless, the idea behind the BCFW recursion relations, that amplitudes can be reconstructed from their analytic behavior, carries over to the loop level, as we’ll now discuss.

## 6 Generalized unitarity and loop amplitudes

Ordinary unitarity is merely the statement that the scattering matrix \(S\) is a unitary matrix, \(S^\dagger S = 1\). Usually we split off a forward-scattering part by writing \(S = 1 + iT\), leading to \((1 - iT^\dagger)(1 + iT) = 1\), or

\[
\text{Disc } T = T^\dagger T, \tag{105}
\]

where \(\text{Disc}(x) = 2 \text{Im}(x)\) is the discontinuity across a branch cut. This equation can be expanded order-by-order in perturbation theory. For example, the four- and five-gluon scattering amplitudes in QCD have the expansions,

\[
\begin{align*}
T_4 &= g^2T_4^{(0)} + g^4T_4^{(1)} + g^6T_4^{(2)} + \ldots, \\
T_5 &= g^3T_5^{(0)} + g^5T_5^{(1)} + g^7T_5^{(2)} + \ldots, \tag{106}
\end{align*}
\]

where \(T_4^{(L)}\) is the \(L\)-loop \(n\)-gluon amplitude. Inserting these expansions into Eq. (105) for the four-point amplitude and collecting the coefficients at order \(g^2\), \(g^4\) and \(g^6\), respectively, we find that,

\[
\begin{align*}
\text{Disc } T_4^{(0)} &= 0, \\
\text{Disc } T_4^{(1)} &= T_4^{(0)}T_4^{(0)}, \\
\text{Disc } T_4^{(2)} &= T_4^{(0)}T_4^{(1)} + T_4^{(1)}T_4^{(0)} + T_5^{(0)}T_5^{(0)}. \tag{108}
\end{align*}
\]

On the right-hand sides of these equations, there is an implicit discrete sum over the types and helicities of the intermediate states which lie between the two \(T\) matrices, and there is a continuous integral over the intermediate-state phase space.
The first equation (generalized to more legs) simply states that tree amplitudes have no branch cuts. The second equation, Eq. (109), states that the discontinuities of one-loop amplitudes are given by the products of tree amplitudes, where the intermediate state always consists of two particles that are re-scattering, the so-called two-particle cuts. The third equation, Eq. (110), states that the discontinuities of two-loop amplitudes are of two types: two-particle cuts where one of the two amplitudes is a one-loop amplitude rather than a tree amplitude, and three-particle cuts involving the product of higher-multiplicity tree amplitudes.

Although there is a lot of information in Eqs. (109) and (110), there are two more observations which lead to even more powerful conclusions. The first observation is that the above unitarity relations are derived assuming real momenta (and positive energies) for both the external states and the intermediate states appearing on the right-hand sides. The intermediate momenta on the right-hand sides can be thought of as particular values of the loop momenta implicit on the left-hand side, momenta that are real and on the particles’ mass shell. Given what we have learned so far about the utility of complex momenta at tree level, it is natural to try to solve the on-shell conditions for the loop momenta for complex momenta as well. Such solutions are referred to as generalized unitarity [30].

Secondly, because unitarity is being applied perturbatively, we might as well make use of other the properties of perturbation theory, namely that a Feynman diagram expansion exists. We don’t need to use the actual values of the Feynman diagrams, but it is very useful to know that such an expansion exists, because we can represent the loop amplitudes as a linear combination of a basic set of Feynman integrals, called master integrals, multiplied by coefficient functions. The idea of the unitarity method [9] is that the information from (generalized) unitarity cuts can be compared with the cuts of this linear combination, in order to determine all of the coefficient functions. If all possible integral coefficients can be determined, then the amplitude itself is completely determined. This approach avoids the need to use dispersion relations to reconstruct full amplitudes from their branch cuts, which is often necessary in the absence of a perturbative expansion.

In the rest of this section, we will sketch a useful hierarchical procedure for determining one-loop amplitudes from generalized unitarity. This method, and variations of it, have been implemented both analytically, and even more powerfully, numerically. The latter implementation has made it possible to compute efficiently one-loop QCD amplitudes of very high multiplicity, far beyond what was imaginable a decade ago. The availability of such loop amplitudes has broken a bottleneck in NLO QCD computations, particularly for processes at hadron colliders such as the LHC, leading to the “NLO revolution.”

6.1 The plastic loop integrand

Before carrying out the loop integration, the integrand of a one-loop amplitude depends on the external momenta \( k_1, k_2, \ldots, k_n \) and on the loop momentum \( \ell \). Just as at tree level, this function can develop poles as the various momenta are continued analytically. Suppose we hold the external momenta fixed and just vary \( \ell \). One kind of singularity that can appear is the ordinary two-particle cut represented by Eq. (109). Let’s first generalize this equation to the case of an \( n \)-gluon one-loop amplitude, and specialize it to the case of a color-ordered loop amplitude \( A_{n}^{l\text{-loop}} \) — the coefficient of the leading-color single-trace color structure discussed in Section 2.

Consider the discontinuity in the channel \( s_{12, m} = (k_1 + k_2 + \cdots + k_m)^2 \), which is illustrated in Fig. 10. The unitarity relation that generalizes Eq. (109) is

\[
\text{Disc}_{s_{12, m}} A_{n}^{l\text{-loop}}(k_1, k_2, \ldots, k_n) = (2\pi)^2 \sum_{h_1} \int \frac{d^D\ell_1}{(2\pi)^D} \delta^{(+)}(\ell_1^\mu) A_{m+2}^{\text{tree}}(-\ell_1^{-h_1}, k_1, \ldots, k_m, \ell_2^{h_2}), \tag{111}
\]

\[
\times \delta^{(+)}(-\ell_2^\mu) A_{n-m+2}^{\text{tree}}(-\ell_2^{-h_2}, k_{m+1}, \ldots, k_n, \ell_1^{h_1}) \tag{112}
\]

where \( \ell_2 = \ell_1 - (k_1 + k_2 + \cdots + k_m) \). The delta function \( \delta^{(+)}(k^\mu) = \Theta(k^0)\delta(k^2) \) enforces that the
intermediate states are on shell with real momenta and positive energies. The sum over intermediate helicities may also include different particle types, for example, both gluons and quarks in an $n$-gluon QCD loop amplitude. The two delta functions reduce the loop momentum integral to an integral over the two-body phase space for on-shell momenta $\ell_1$ and $-\ell_2$.

Another way of stating Eq. (111), which allows us to generalize it, is that for a given set of external momenta $k_i$, there is a family of loop momenta $\ell \equiv \ell_1$ that solve the dual constraints $\ell_1^2 = \ell_2^2 = 0$. On this solution set the loop integrand, which can be pictured as the annular blob shown in fig. 10, factorizes into the product of two tree amplitudes, in much the same way that a tree amplitude factorizes on a single multi-particle pole, Eq. (69).

In this picture of the plastic loop integrand, we need not impose positivity of the energies of the intermediate states, and the loop momenta can even be complex. This opens up the possibility of more general solutions, where more than two lines are cut. If we think of the loop momentum $\ell^\mu$ as four-dimensional, then for generic kinematics we can cut not just two lines, but up to four. The reason the maximum is four is that each cut imposes a new equation of the form $(\ell - K_i)^2 = 0$ for some combination of external momenta $K_i$. At four cuts the number of equations equals the number of unknowns — the four components of $\ell^\mu$. Hence a fifth cut condition is impossible to satisfy (unless the

$$\begin{align*}
\frac{i}{\ell_1^2} A_{m+2}^{\text{tree}}(-\ell_{h_1}^1, k_1, \ldots, k_m, \ell_{h_2}^2) \frac{i}{\ell_2^2} A_{n-m+2}^{\text{tree}}(-\ell_{h_2}^2, k_{m+1}, \ldots, k_n, \ell_{h_1}^1),
\end{align*}$$

(113)

in much the same way that a tree amplitude factorizes on a single multi-particle pole, Eq. (69).
$$A_n^{1\text{-loop}} = \sum_i d_i + \sum_i c_i + \sum_i b_i + R_n + O(\epsilon)$$

Fig. 12: Decomposition of a generic one-loop amplitude $A_n^{1\text{-loop}}$ into basis integrals multiplied by kinematical coefficients: scalar box integrals with coefficients $d_i$, scalar triangles with coefficients $c_i$, scalar bubbles with coefficients $b_i$, and the rational part $R_n$. The dots between the external lines indicate that one or several external legs may emanate from each vertex. If there are massive internal propagators, then tadpole integrals also appear; in the massless case such integrals vanish.

kinematical configuration of the external momenta is an exceptional, degenerate one). Figure 11 shows how the quadruple cut of a generic one-loop integrand squeezes it at four locations, so that it becomes proportional to the product of four tree amplitudes. Two of the momenta of each tree amplitude are identified with the cut loop momenta, denoted by $\ell_1, \ell_2, \ell_3, \ell_4$, and the rest are drawn from the external momenta for the loop amplitude.

6.2 The quadruple cut

The quadruple cut [31] is special because the solution set is discrete. Let’s write the four cut loop momenta as

$$\ell_1, \quad \ell_2 = \ell_1 - K_1, \quad \ell_3 = \ell_2 - K_2, \quad \ell_4 = \ell_3 - K_3 = \ell_1 + K_4,$$

where the $K_i$ are sums of the $n$ external momenta satisfying $K_1 + K_2 + K_3 + K_4 = 0$. From Fig. 11 it is clear that the $K_i$ correspond to some partition of the $n$ cyclicly ordered momenta into four contiguous sets. We can rewrite the four quadratic cut conditions,

$$\ell_1^2 = \ell_2^2 = \ell_3^2 = \ell_4^2 = 0, \quad (115)$$

by taking the differences $\ell_i^2 - \ell_{i+1}^2 = 0$, so that three of the conditions are linear,

$$\ell_1^2 = 0, \quad 2\ell_1 \cdot K_1 = K_1^2, \quad 2\ell_2 \cdot K_2 = K_2^2, \quad 2\ell_3 \cdot K_3 = K_3^2. \quad (116)$$

Because the three linear equations can be solved uniquely, we generically expect two discrete solutions for the loop momentum $\ell_1$, denoted by $\ell_1^\pm$. The other three quantities $\ell_i^\pm$ are uniquely determined from $\ell_1^\pm$ by shifting it by the appropriate external momenta.

What information does the quadruple cut reveal? To answer this question, we rely on a systematic decomposition of the one-loop amplitude for an arbitrary $n$-point amplitude, which is shown diagrammatically in Fig. 12. The amplitude can be written as a linear combination of certain basis integrals, multiplied by kinematical coefficients. The only loop integrals that appear are scalar integrals with four, three and two internal propagator lines, which are usually called box, triangle and bubble integrals, respectively. They are given in dimensional regularization, with $D = 4 - 2\epsilon$, by

$$I_4(K_1, K_2, K_3, K_4) = \mu^{2\epsilon} \int \frac{d^{4-2\epsilon} \ell}{(2\pi)^{4-2\epsilon}} \frac{1}{\ell^2(\ell - K_1)^2(\ell - K_1 - K_2)^2(\ell + K_4)^2}, \quad (117)$$

57
The integrals are very easy to evaluate, either analytically or in an automated code, and they are numerically very stable. Here the external momenta \( \{ K_1, K_2, K_3 \} \) are the sums of external momenta emanating from each corner. The coefficients of these integrals are \( d_i, c_i \) and \( b_i \), where \( i \) labels all the inequivalent partitions of the \( n \) external momenta into 4, 3 and 2 sets, respectively. There is also a rational part \( R_n \), which cannot be detected using cuts with four-dimensional cut loop momenta; we will return to this contribution later.

The decomposition in Fig. 12 holds in dimensional regularization, assuming that the external (observable) momenta are all four-dimensional, and neglecting the \( O(\epsilon) \) terms. It also requires the internal propagators to be massless; if there are internal propagators for massive particles, then tadpole (one-propagator) integrals will also appear. The result seems remarkable at first sight, since one-loop Feynman diagrams with five or more external legs attached to the loop will generically appear, and these diagrams would seem likely to generate pentagon and higher-point integrals. However, it is possible to systematically reduce such integrals down to linear combinations of scalar boxes, triangles and bubble integrals [32–34].

The reduction formulas are fairly technical, but here we don’t need to know the formulas, just that the reduction is possible. Heuristically, the reason it is possible to avoid all pentagon and higher-point integrals is the same reason that there is no quintuple cut when the loop momentum is in four dimensions: there are more equations in the quintuple cut conditions than there are unknowns. If the scalar pentagon integral had a quintuple cut, it would not be possible to reduce it to a linear combination of box integrals. The fact that it can be done [32] exploits the four-dimensionality of the loop momenta to expand the loop momenta in terms of the four linearly-independent external momenta of the pentagon. In dimensional regularization, the relation of ref. [32] has a correction term [33], and the pentagon integral has a quintuple cut, because the loop momentum is no longer four-dimensional. However, because of the “small” volume of the extra \(-2\epsilon\) dimensions, the correction term is of \( O(\epsilon) \).

Returning to the quadruple cut, we see that a second special feature of it is that only one of the integrals in Fig. 12 survives, for a given quadruple cut. First of all, none of the triangle and bubble terms can survive, because those integrals do not even have four propagators available to cut. There are many possible box integrals, for a large number of external legs, but each one box integral is in one-to-one correspondence with a different quadruple cut; both are characterized by the same partition of the cyclicly ordered momenta into four contiguous sets, or clusters. The momentum flowing out at each corner of the box must match the cluster momenta \( \{ K_1, K_2, K_3, K_4 \} \) corresponding to the quadruple cut (115). For this solution, we match the left- and right-hand sides of Fig. 12 and learn [31] that

\[
d_i = \frac{1}{2}(d_i^+ + d_i^-),
\]

where the superscripts \( \pm \) refer to the two discrete solutions for the loop momentum, and \( d_i^\pm \) are given by the product of four tree amplitudes, as in Fig. 11,

\[
d_i^\pm = A_1^{\text{tree}}(\ell^\pm)A_2^{\text{tree}}(\ell^\pm)A_3^{\text{tree}}(\ell^\pm)A_4^{\text{tree}}(\ell^\pm),
\]

with

\[
A_i^{\text{tree}}(\ell) \equiv A^{\text{tree}}(-\ell, k_1^{(i)}, \ldots, k_p^{(i)}, \ell_{i+1}).
\]

Here the external momenta \( \{ k_1^{(i)}, \ldots, k_p^{(i)} \} \) are the elements of the cluster \( K_i \), \( i = 1, 2, 3, 4 \), i.e. \( \sum_{j=1}^{p_i} k_j^{(i)} = K_i \). These formulae are very easy to evaluate, either analytically or in an automated code, and they are numerically very stable.
A brief Introduction to Modern Amplitude Methods

Fig. 13: The left quadruple cut shows that the coefficients of all four-mass box integrals vanish for one-loop NMHV amplitudes. The right quadruple cut shows that the three-mass box coefficients do not vanish.

It’s possible to solve analytically for the cut loop momenta $\ell_i^\pm$ for generic values of the $K_i$; the solution involves a quadratic formula [31]. If just one of the external legs is massless, however, say $K_1 = k_1$, then the solutions collapse to a simpler form [35, 36]:

\[
\begin{align*}
(\ell_1^\pm)^\mu &= \frac{\langle 1^\pm | K_2 K_3 K_4 | 1^\pm \rangle}{2 \langle 1^\pm | K_2 K_4 | 1^\pm \rangle}, \\
(\ell_2^\pm)^\mu &= -\frac{\langle 1^\pm | K_2 K_3 | 1^\pm \rangle}{2 \langle 1^\pm | K_2 K_4 | 1^\pm \rangle}, \\
(\ell_3^\pm)^\mu &= \frac{\langle 1^\pm | K_2 K_3 K_4 | 1^\pm \rangle}{2 \langle 1^\pm | K_2 K_4 | 1^\pm \rangle}, \\
(\ell_4^\pm)^\mu &= -\frac{\langle 1^\pm | K_2 K_3 K_4 | 1^\pm \rangle}{2 \langle 1^\pm | K_2 K_4 | 1^\pm \rangle}.
\end{align*}
\]

(123)

It’s easy to see that Eq. (115) is satisfied by Eq. (123); that is, each of the four vectors $(\ell_i^\pm)^\mu$ squares to zero. For example, the evaluation of $(\ell_1^\pm)^\mu (\ell_1^\pm)_\mu$ proceeds using the Fierz identity and is proportional to $\langle 1 1 \rangle = 0$. The corresponding algebra for $(\ell_2^\pm)^2$ involves $\langle 1^\pm | K_2 K_2 | 1^\pm \rangle = K_2^2 \langle 1 1 \rangle = 0$.

We also have to show that momentum conservation is satisfied, namely,

\[
\ell_2 - \ell_3 = K_2, \quad \ell_3 - \ell_4 = K_3, \quad \ell_4 - \ell_1 = K_4. \tag{124}
\]

The first equation is

\[
(\ell_2^\pm - \ell_3^\pm)^\mu = -\frac{\langle 1^\pm | [\gamma^\mu, K_2] K_3 K_4 | 1^\pm \rangle}{2 \langle 1^\pm | K_2 K_4 | 1^\pm \rangle} = -K_2^\mu \frac{\langle 1^\pm | (K_1 - K_2 - K_4) K_4 | 1^\pm \rangle}{\langle 1^\pm | K_2 K_4 | 1^\pm \rangle} = K_2^\mu, \tag{125}
\]

and the other equations work the same way.

Shortly, we will compute an explicit example of a nontrivial, nonzero coefficient of a box integral using the quadruple cut. However, it’s worth noting first that many box coefficients for massless QCD amplitudes vanish identically. In fact, the vanishing of large sets of box coefficients can be established simply by counting negative helicities. Consider, for example, the one-loop NMHV amplitude in massless QCD whose quadruple cut is shown on the left side of Fig. 13. This quadruple cut can be used to compute the coefficient of a four-mass box integral. We call it a four-mass box because the momentum $K_i$ flowing out at each corner is the sum of at least two massless external particle momenta; hence $K_i$ is a massive four-vector. (In contrast, the right side of Fig. 13 shows a quadruple cut for a three-mass box integral, because the lower right tree amplitude emits a single external momentum $m$.)

We denote negative-helicity legs by red lines and an explicit (−) in the figure. The external black lines are all positive helicity. The upper left tree amplitude in the example has no external negative helicities. Because tree amplitudes with 0 or 1 negative helicity vanish, according to Eq. (67), the two internal (cut) lines emanating from this upper left blob must carry negative helicity. On the opposite side of their respective cuts, they carry positive helicity. If the lower left and upper right tree amplitudes have one negative external helicity, as shown, then they must each send a negative helicity state toward the purple blob. This tree amplitude carries the third external negative helicity, but no other negative helicity emanates from it, so it vanishes, causing the vanishing of the corresponding four-mass box coefficient.
We gave this argument specifically for the case that all three negative-helicity particles were emitted from different corners of the box. It’s easy to see that the vanishing does not actually depend on where the negative helicities are located. It’s simply a reflection of the fact that there are four tree amplitudes, all with more than three legs, so there must be at least $4 \times 2 = 8$ negative helicities among the external and cut legs. However, each cut has exactly one negative helicity, and there are three negative external helicities, for a total of $4 + 3 = 7$. Since $7 < 8$, the NMHV four-mass box coefficients always vanish. This counting argument fails as soon as one of the corner momenta becomes massless, as is appropriate for the three-mass cut shown on the right side of Fig. 13. With the right (second) type of complex kinematics discussed in Section 4.5, the three-point tree amplitude with helicity configuration $(++-)$ is nonvanishing, as shown in the figure. Hence this three-mass box coefficient is nonvanishing. There is a single quadruple-cut helicity configuration and a single choice of sign for the kinematical configuration (123) that contributes in the particular case shown.

Using the same counting argument, we can see that one-loop MHV amplitudes, with two external negative helicities, contain neither four-mass, nor three-mass, box integrals. The two-mass box integrals can be divided into two types, “easy”, in which the two massive corners are diagonally opposite, and “hard”, in which they are adjacent to each other. One can show that the hard two-mass boxes always vanish as well. (This proof can be done with the help of a triple cut which puts the two massless corners into one of the three trees. Then the counting of negative helicities is analogous to the four-mass NMHV example, except that one needs $3 \times 2 = 6$ negative helicities, and one has only $3 + 2 = 5$ available.)

As an aside, consider the one-loop amplitudes of the form $A_{4\text{loop}}^\pm(1^\pm, 2^+, 3^+, 4^+, 5^+)$, for which the corresponding tree amplitudes vanished according to Eq. (67). A similar counting exercise shows that they have no cuts at all: no quadruple, triple, or ordinary two-particle cuts. They are nonvanishing (at least in a non-supersymmetric theory like QCD), but they are forced to be purely rational functions of the external kinematics [37].

### 6.3 A five-point MHV box example

In the remainder of this section, we will compute one of the box coefficients for the five-gluon QCD amplitude $A_{5\text{loop}}^\pm(1^-, 2^-, 3^+, 4^+, 5^+)$, the one in which the two negative helicity legs, 1 and 2, are clustered into a massive leg (as also reviewed in ref. [10]). The quadruple cut for this box coefficient is shown in Fig. 14. Inspecting the figure, starting with the lower-left tree amplitude, it is clear that there is a unique assignment of internal helicities. Also, this assignment of helicities forbids quarks (or scalars) from propagating in the loop; the tree amplitudes for two spin 1/2 fermions (or two scalars) and two identical helicity gluons vanish (see Eq. (71) for the fermion case). Therefore this box coefficient receives contributions only from the gluon loop, and is the same in QCD as in gauge theories with different matter content (such as $\mathcal{N} = 4$ super-Yang-Mills theory).

Now that we have identified which four tree amplitudes are to be multiplied together, the next task is to determine the cut loop momentum. In particular, let’s work out $\ell_4$, the loop momentum just before the massless external leg 4. We can use Eq. (123), but since leg 1 was massless there, we should relabel the momenta in that equation according to:

$$\ell_1^\pm \rightarrow \ell_4^\pm, \quad k_1 \rightarrow k_4, \quad K_2 \rightarrow K_5, \quad K_3 \rightarrow k_1 + k_2, \quad K_4 \rightarrow k_3. \quad (126)$$

Then the first equation in (123) becomes,

$$\ell_4^\pm = \ell_4^\pm \quad (4^\pm 5(1 + 2)3\gamma^\mu | 4^\pm \rangle = \frac{\langle 4^\pm | 543\gamma^\mu | 4^\pm \rangle}{2 \langle 4^\pm | 53 | 4^\pm \rangle} = -\frac{\langle 5^\pm | 43\gamma^\mu | 4^\pm \rangle}{2 \langle 5^\pm | 3 | 4^\pm \rangle}. \quad (127)$$

Which sign should we use? The sign is dictated by the helicity assignments in the three-point amplitudes. Because the upper-right tree is of type $(-+)$, and is constructed from right-handed spinors, the three left-handed spinors should be proportional. In particular, $\lambda_4 \propto \lambda_4$, which tells us that we should take
the lower sign in Eq. (127), so that
\[ \ell_4^\mu = (\ell_4^-)^\mu = \frac{1}{2} \langle 45 \rangle^3 \langle 3^- | \gamma^\mu | 4^- \rangle. \] (128)

Now we can multiply together the four tree amplitudes, and use Eqs. (120) and (121) (with \( d_t^+ = 0 \)) to get for the “\((12)^n\)” box coefficient,
\[ d_{(12)} = \frac{1}{2} A_4^{\text{tree}} (-\ell_1^\mu, \ell_2^\mu, \ell_3^\mu, \ell_4^\mu) \]
\[ = \frac{1}{2} \langle 12 \rangle^3 \langle 3 \ell_4 \rangle^3 \langle 5 \ell_5 \rangle^3 \langle (-\ell_4^-) \ell_5^- \rangle^3 \]
\[ = \frac{1}{2} \langle 2 \ell_3 \rangle \langle 3 \ell_3 \rangle \langle (-\ell_1^-) 1 \rangle \langle (\ell_4^-) 3 \rangle \langle (\ell_5^-) 5 \rangle \langle (\ell_4^-) 4 \rangle \langle (\ell_5^-) 5 \rangle \langle (\ell_4^-) 3 \rangle \langle (\ell_5^-) 5 \rangle \]
\[ = -\frac{1}{2} \langle 2 \ell_3 \rangle \langle 3 \ell_3 \rangle \langle (-\ell_1^-) \ell_4^- \ell_5^- \rangle \langle 4^- | \ell_4^- \ell_5^- \ell_1^- \rangle \langle 5^- | \ell_4^- \ell_5^- \ell_1^- \rangle. \] (129)

To get to the last step in eq. (129), we combined spinor products into longer strings using the replacement \( |\ell_i\rangle |\ell_i\rangle \rightarrow \theta_i \), but we did not need to use any other properties of the \( \ell_i \). In the next step it is convenient to use momentum conservation, \( \ell_1 = \ell_4 - k_1 - k_5, \ell_3 = \ell_4 + k_3 \) and \( \ell_5 = \ell_4 - k_4 \), as well as \( \ell_4^2 = 0 \), to replace,
\[ \langle 3^- | \ell_4 \ell_5 \rangle \rightarrow \langle 4^- | \ell_4 | 3^- \rangle \langle 45 \rangle, \]
\[ \langle 2^- | \ell_3 \ell_4 \rangle \rightarrow \langle 2^- | \ell_4 | 3^- \rangle \]
\[ \langle 4^- | \ell_4 \ell_3 \ell_1 \rangle \rightarrow \langle 4^- | \ell_4 \ell_3 \rangle \langle 45 \rangle \]
\[ \langle 1^- | \ell_1 \ell_4 \rangle \rightarrow \langle 5^- | \ell_4 | 4^- \rangle. \] (130) (131) (132) (133)

In Eq. (132) we also used the fact that \( \langle 3 \ell_4 \rangle = 0 \), given that both \( \ell_4 \) and \( k_3 \) emanate from a \((++-\)) three-point amplitude.

Making these replacements in Eq. (129), and then Fierzing in \( \ell_4^\mu \propto \langle 3^- | \gamma^\mu | 4^- \rangle \) from Eq. (128), gives,
\[ d_{(12)} = \frac{1}{2} \langle 12 \rangle^3 \langle 4^- | \ell_4 | 3^- \rangle^2 \langle 45 \rangle^3 \]
\[ = -\frac{1}{2} \langle 23 \rangle \langle 34 \rangle \langle 45 \rangle \langle 51 \rangle \]

Fig. 14: The quadruple cut for one of the box coefficients for the five-gluon amplitude with helicity configuration (---+++).
\[ I^{(12)}_4 = \frac{-2i \Gamma(1 + \epsilon) \Gamma^2(1 - \epsilon)}{(4\pi)^2 - \epsilon} \frac{c_T}{\Gamma(1 - 2\epsilon)} \] (136)

Interestingly, the result (134) is proportional to the tree amplitude. The coefficients of the four other box integrals (labeled (23), (34), (45) and (51)) also have only gluonic contributions for this helicity choice, and their coefficients turn out to be given by cyclic permutations of Eq. (134). Hence we have for the gluonic contribution to the one-loop amplitude,

\[ A_5^{\text{1-loop}}(1^-, 2^-, 3^+, 4^+, 5^+) = A_5^{\text{tree}}(1^-, 2^-, 3^+, 4^+, 5^+) c_T \left\{ -\frac{1}{\epsilon^2} \left( \frac{\mu^2}{-s_{34}} \right) + \left( \frac{\mu^2}{-s_{45}} \right) - \left( \frac{\mu^2}{-s_{12}} \right) \right\} + \text{cyclic permutations} \]

(137)

If we were computing the amplitude in \( \mathcal{N} = 4 \) super-Yang-Mills theory, we would be done at this point. One can show that the triangles, bubbles and rational parts all vanish in this theory [9]. In the case of QCD, there is more work to do. In the next subsection we sketch a method [35, 38] for determining the triangle coefficients.

### 6.4 Triangle coefficients

By analogy, we expect the triangle coefficients to be determined by the triple cut shown in Fig. 15(a), and the bubble coefficients by the double cut shown in Fig. 15(b). The solution to the three equations defining the triple cut,

\[ \ell_1^3(t) = \ell_2^3(t) = \ell_3^3(t) = 0, \] (138)

depends on a single complex parameter \( t \). However, the triple cut generically also receives contributions from the box integral terms in Fig. 12. The box contributions have to be removed before identifying the coefficient of a given scalar triangle integral. Take any one of the three tree amplitudes in Fig. 15(a), and imagine pinching that blob until it splits into two, exposing another loop propagator. This corner of the triple-cut phase space has the form of a box integral contribution. The pinching imposed a fourth cut condition, which has discrete solutions, so it must only occur at discrete values of \( t \), say \( t_i \) where \( i \) labels the different quadruple cuts that sit “above” the given triple cut, and \( \sigma = \pm \) labels the two possible discrete solutions.
The generic form of the triple cut is
\[
C_3(t) = A^{tree}_{(1)}(t_1, k_1, \ldots, k_{p_1}, t_2) + \sum_{\sigma=\pm} A^{tree}_{(2)}(-t_2, k_{p_1+1}, \ldots, k_{p_2}, t_3) + A^{tree}_{(3)}(-t_3, k_{p_2+1}, \ldots, k_n, t_1)
\]
\[
= T_3(t) + \sum_{\sigma=\pm} \sum_{i} d_{\sigma i}^{\ell_i} \xi_{\sigma i} (t - t_{\sigma i}),
\]
(139)

where \(d_{\sigma i}^{\ell_i}\) are the previously computed box coefficients (121), and \(T_3(t)\) is the triple cut “cleaned” of all singularities at finite \(t\).

The pole locations \(t_{\sigma i}\) and the residue factors \(\xi_{\sigma i}\) do not depend on the amplitude being calculated, but only on the kinematics of the relevant triple and quadruple cuts. They can be computed from the solution for \(\ell_i(t)\). For massless internal particles, the solution of Eq. (138) is [38–40]
\[
\ell_1^{\alpha}(t) = \tilde{K}_1^{\alpha} + K_2^{\alpha} + \frac{t}{2} (\tilde{K}_1^{-} |\gamma_{\mu}^{\alpha}| \tilde{K}_3^{-}) + \frac{1}{2t} (\tilde{K}_3^{-} |\gamma_{\mu}^{\alpha}| \tilde{K}_1^{-}) ,
\]
(140)

and, using momentum conservation, \(\ell_3(t) = \ell_1(t) - K_1, \ell_3(t) = \ell_1(t) + K_3\). Here we have introduced a pair of massless auxiliary vectors \(\tilde{K}_1^{\mu}\) and \(\tilde{K}_3^{\mu}\), constructed from \(K_1\) and \(K_3\),
\[
\tilde{K}_1^{\mu} = \gamma_{\alpha}^{\mu} K_1^{\alpha} + \frac{1}{\gamma^2 - S_1 S_3}, \quad \tilde{K}_3^{\mu} = -\gamma_{\alpha'}^{\mu} K_3^{\alpha} + \frac{1}{\gamma^2 - S_1 S_3},
\]
(141)

where \(S_1 = K_1^\alpha K_1^\alpha, S_3 = K_3^\alpha K_3^\alpha, \) and
\[
\alpha = \frac{S_3 (S_1 - \gamma)}{S_1 S_3 - \gamma^2}, \quad \alpha' = \frac{S_1 (S_3 - \gamma)}{S_1 S_3 - \gamma^2}, \quad \gamma = \gamma_{\pm} = -K_1 \cdot K_3 \pm \sqrt{\Delta},
\]
(142)

with
\[
\Delta = (K_1 \cdot K_3)^2 - K_1^2 K_3^2.
\]
(143)

The coefficient of the scalar triangle integral is the “\(t\) independent” part of the triple cut. To be more precise, the quantity \(T_3(t)\) has no singularities at finite values of \(t\) because they are all accounted for by the box contributions shown explicitly in Eq. (139). Because this quantity has singularities only at \(t = 0\) and \(t = \infty\), it can be represented as,
\[
T_3(t) = \sum_{k=-p}^{p} c_k t^k.
\]
(144)
The desired quantity, the triangle coefficient, is \( c_0 \). The other terms correspond to tensor triangle integrals that integrate to zero ("spurious terms" in the language of OPP [40]). For renormalizable theories, there are at most three loop momenta in the numerator of triangle integrals, and one can take \( p = 3 \).

One way to isolate \( c_0 \) is from the \( \ell^0 \) term in the large \( t \) limit of \( T_3(t) \), or of \( C_3(t) \) itself, since the box contributions go to zero in this limit. This is an effective method for determining \( c_0 \) analytically [38]. For an automated implementation, the \( \ell^0 \) term is usually subleading as \( t \to \infty \), making it difficult to extract numerically. Instead one can work at finite \( t \), and extract \( c_0 \) (and the other \( c_k \) coefficients) out of the finite sum in Eq. (144) by using the discrete Fourier projection,

\[
c_k = \frac{1}{2p + 1} \sum_{j=-p}^{p} \left[ t_0 e^{2\pi ij/(2p+1)} \right]^{-k} T_3 \left( t_0 e^{2\pi ij/(2p+1)} \right),
\]

for some choice of \( t_0 \). This approach is very stable numerically [35].

The other \( c_k \) coefficients are actually needed in the next step, the determination of the bubble coefficients. The double cut depends on two complex parameters. It has singularities corresponding to both triple cuts and quadruple cuts, which can be "cleaned" in a fashion analogous to Eq. (139), using the previously computed box and triangle information. Because the triple cut depends on a complex parameter, all of the \( c_k \) coefficients are required to characterize it. After cleaning the double cut, a double discrete sum analogous to Eq. (145) can be used to extract the bubble coefficient. For real cut momenta, the two parameters of the double cut have a simple physical interpretation: they are just the angles \( \theta, \phi \) of one of the two intermediate states, in the center of mass frame for the channel being cut. The double discrete sum essentially performs a spherical harmonic expansion (it is slightly different because the intermediate momenta can be treated as complex).

The hierarchical determination of the “cut-constructible” parts of one-loop amplitudes described here [35] is quite similar to the OPP method [40] and to the method described in Ref. [41], all of which have been implemented in an automated fashion.

### 6.5 The rational part

The last remaining part of the amplitude is the rational part \( R_n \). This component cannot be detected by any unitarity cut in which the cut loop momentum are confined to four dimensions. We have implicitly been assuming throughout this section that the \( \ell_{ij} \) are four-dimensional. This assumption was very convenient because it allowed us to label the states with four-dimensional helicities, and use all the vanishing relations for the tree amplitudes that enter the four-dimensional cuts. One way to determine the rational part, called \( D \)-dimensional unitarity [18,42,43], is to let the cut momenta have extra-dimensional components, thinking of the \( \epsilon \) in \( D = 4 - 2\epsilon \) as a negative number. In this approach, there are also nonvanishing quintuple cuts. There are no hexagon cuts because at one loop, all extra-dimensional components of the loop momentum are equivalent; they might as well point in a single, fifth dimension. So there are five components of the loop momentum that can be constrained by generalized cuts. The same kind of hierarchical, automated approach described above can be applied to the \( D \)-dimensional case [44]. In this case, one does not need to determine every extra-dimensional term in the loop integrand; the measure factor is \( d^{2\ell} \ell \), leading to an integral of \( O(\epsilon) \), unless there are enough factors of the extra-dimensional components, denoted by \( \ell^2_{(\epsilon)} \equiv \mu^2 \), in the numerator of the loop integrand to generate a compensating factor of \( 1/\epsilon \). For more details on this method, see the review [13].

A second method for computing the rational part is to apply a BCFW shift to the integrated loop amplitude. This approach can be implemented analytically [10], and numerically [35]. Here we just mention a few salient points. When a complex \( z \)-dependent shift is applied to a tree amplitude, as in Section 5, the result is a meromorphic function of \( z \), where the poles correspond to factorization of the tree amplitude into two lower-point amplitudes. When the same shift is applied to a loop amplitude, branch cuts in \( z \) are generated, from the logarithms and dilogarithms appearing in the scalar integrals.
There are also poles, whose origin from amplitude factorization is similar to the tree-level case. The branch cuts would complicate an analysis of the poles. However, if we have already computed the cut part $C_n$, we can consider shifting only the rational part, $R_n = A_n - C_n \to R_n(z)$.

The function $R_n(z)$ is meromorphic, so we can contemplate computing $R_n(0)$ from Cauchy’s theorem, using an equation analogous to eq. (80), if we know all of its poles and residues. However, $R_n(z)$ has two different types of poles. The physical poles are the ones that appear in $A_n(z)$, and their residues can be computed from factorization in a similar fashion to tree level. There is a second set of spurious poles. These poles are not poles of $A_n(z)$. They come from singularities in kinematical regions where $A_n$ is non-singular, but $C_n$ and $R_n$ separately diverge. (One example of such a region is where $\langle 2^\mu | (6 + 1) | 5^- \rangle \to 0$; see section 5.3.) Because $A_n(z)$ has no poles, the spurious-pole residues in $R_n(z)$ must be the negative those in $C_n(z)$. Because the cut part is known and the locations of all the spurious poles are known, the residues of $C_n(z)$ are straightforward to compute. For more details on this method, see the review [10].

Within the OPP method [40], the rational part is given by a sum of two terms, called $R_1$ and $R_2$. The $R_1$ part is obtained as a byproduct of the computation of the cut part, by taking into account the extra-dimensional $\mu^2$ dependence appearing in the propagator denominators of the dimensionally-regulated loop integrand [45]. The remaining $R_2$ terms come from $\mu^2$ dependence in the numerator of the loop integrand. As in the $D$-dimensional unitarity method, only a limited set of terms have enough factors of $\mu^2$ in the numerator to produce a nonzero rational term. For renormalizable theories, these contributions can be computed for all processes, in terms of a relatively small number of effective two-, three- and four-point vertices [45, 46].

These new, efficient methods have enabled the construction of a variety of automated computer programs for generating one-loop amplitudes, including CUTTOOLS [47], BLACKHAT [35], ROCKET [48], SAMURAI [49], NGLUON [50], MADLOOP [51], HELAC-NLO [52], GoSAM [53], OPEN LOOPS [54] and RECOLA [55].

For NLO QCD corrections to collider processes, it is also necessary to consider tree-level processes with one additional parton radiated into the final state, and integrate their cross section over a phase space that contains the soft and collinear singularities discussed in section 4. A variety of efficient, automated methods have been developed recently for performing these phase-space integrals, based on the methods originally developed in refs. [56, 57]. In combination with the one-loop methods sketched here, these methods have led to a variety of NLO QCD results for LHC processes with four, five and even six objects (electroweak particles or jets) in the final state. They have opened up a new avenue for precision theory at hadron colliders, which has proved to be very important for gaining quantitative control over important Standard Model backgrounds, as well as for performing detailed experimental studies of QCD dynamics.

7 Conclusions

In these notes, we have only scratched the surface of modern techniques for computing scattering amplitudes. We covered the general formalism and factorization properties of helicity amplitudes, explored tree-level analyticity and the BCFW recursion relation, and described some of the techniques for using generalized unitarity at one loop. Numerous additional details are required in order to assemble full one loop QCD amplitudes, many of which are discussed in other reviews [10–12], and in particular the comprehensive review [13].

We did not touch on multi-loop scattering amplitudes at all, but this is an exceedingly rich subject. Amplitudes in $\mathcal{N} = 4$ super-Yang-Mills theory — QCD’s maximally supersymmetric cousin — have been computed using similar ideas, through many loops and for many external legs. Remarkable properties have been found, leading to new approaches. For more in this direction, as well as applications to supergravity, the reader can consult the very recent, authoritative review [14].

65
The multi-loop applications of unitarity-based methods to QCD are still in their infancy, but they are being developed very rapidly now. For the simplest $2 \rightarrow 2$ processes, the principles of generalized unitarity were applied a while ago [58–60], but not in a way that could be automatically extended to more complicated processes. The latter direction has seen important recent progress [61–64], but there is still a ways to go before two-loop QCD amplitudes for generic $2 \rightarrow 3$ processes will be available. A large part of the problem is not just determining the loop integrand, but evaluating all the loop integrals.

I hope that some of you who have made it this far will be encouraged to explore further, and indeed to push the boundaries of our knowledge about scattering amplitudes and their applications to collider physics as well as other problems.

Acknowledgments

I am grateful to Christophe Grojean, Martijn Mulders, Maria Spiropolu, Bogdan Dobrescu and Iain Stewart for the invitations to give these two sets of lectures and for the encouragement to prepare these notes. I also thank the students at both schools for their enthusiasm and excellent questions. Finally, I thank Zvi Bern, David Kosower and my other colleagues for many enjoyable collaborations on related topics. The figures in this contribution were generated using Jaxodraw [65], based on Axodraw [66]. This work was supported by the US Department of Energy under contract DE-AC02-76SF00515.

References


[arXiv:0903.0356 [hep-ph]];
[arXiv:1110.1499 [hep-ph]].
[hep-ph/9605323].
S. Badger, H. Frellesvig and Y. Zhang, A two-loop five-gluon helicity amplitude in QCD,
arXiv:1310.1051 [hep-ph].
[arXiv:1206.4180 [hep-ph]].
Flavor Physics and CP Violation

G. Isidori
INFN, Laboratori Nazionali di Frascati, Frascati, Italy
CERN, Geneva, Switzerland

Abstract
These lectures cover the following topics: 1) flavor physics within the Standard Model, 2) phenomenology of $B$ and $D$ decays, 3) flavor physics beyond the Standard Model.

Introduction
According to the Standard Model (SM) of fundamental interactions the basic constituents of matter, and their interactions, are described as excitations of fermionic fields (spin-1/2 particles) interacting with there different sets of gauge fields (whose excitations correspond to spin-1 particles) associated to the strong, weak, and electromagnetic interactions. The spin-1/2 particles can de grouped into three families, or flavors, each containing two quarks (charged under strong interactions) and two leptons (neutral under strong interactions). The four fermions within each family have different combinations of strong, weak, and electromagnetic charges, that determine completely their fundamental interactions but for gravity. Ordinary matter consists essentially of particles of the first family, namely the up and down quarks (strongly bounded inside protons and neutrons), the electrons (that forms the atoms), and the electron-neutrinos (abundantly produced by the fusion reactions occurring inside the stars). As far as we know, quarks and leptons of the second and third family are identical copies of those in the first family but for different, heavier, masses. The heavy quarks and charged leptons are unstable states that can be produced in high-energy collisions and that decay very fast (via weak interactions) into particles of the first family. Why we have three almost identical replica of quarks and leptons, and which is the origin of their different masses, is one of the big open questions in fundamental physics.

In the limit of unbroken electroweak symmetry none of the basic constituent of matter could have a non-vanishing mass. The problem of quark and lepton masses is therefore intimately related to the other big open question in particle physics: which is the mechanism behind the breaking of the electroweak symmetry, or which is the mechanism responsible for the non-vanishing masses of the weak force carriers (the $W$ and $Z$ bosons). Within the SM these two problems are both addressed by the Higgs mechanism: the masses of quarks and leptons, as well as the masses of $W$ and $Z$ bosons, are the result of the interaction of these basic fields (both matter constituents and force carriers) with a new type of field, the Higgs scalar field, whose ground state breaks spontaneously the electroweak symmetry.

The recent observation by the ATLAS and CMS experiments of a new state compatible with the properties of the Higgs boson (or the spin-0 excitation of the Higgs field) has significantly reinforced the evidences in favor of the Higgs mechanisms and the validity of the SM. However, we have also clear indications that this theory is not complete: the phenomenon of neutrino oscillations and the evidence for dark matter cannot be explained within the SM. The SM is also affected by a serious theoretical problem because of the instability of the Higgs sector under quantum corrections. We have not yet enough information to unambiguously determine how this theory should be extended; however, most realistic proposals point toward the existence of new degrees of freedom in the TeV range, possibly accessible at the high-$p_T$ experiments at the LHC.

The description of quark and lepton masses in terms of the Higgs mechanism is particularly unsatisfactory since the corresponding interactions is not controlled by any symmetry principle, contrary to all other known interactions, resulting in a large number of free parameters. Beside determining quark masses, the interaction of the quarks with the Higgs is responsible for the peculiar pattern of mixing of...
the various families of quarks under weak interactions, and the corresponding hierarchy in the various
decay modes of the heavier quarks into the lighter ones. In particular, the interplay of weak and Higgs
interactions implies that processes with a change of flavor mediated by a neutral current (FCNC pro-
cesses) can occur only at higher orders in the electroweak interactions and are strongly suppressed. This
strong suppression make FCNC processes natural candidates to search for physics beyond the SM: if the
new degrees of freedom do not have the same flavor structure of the quark-Higgs interaction present in
the SM, then they could contribute to FCNC processes comparably to the SM amplitudes even if their
masses are well above the electroweak scale, resulting in sizable deviations from the SM predictions for
these rare processes.

In the last few years the mechanism of quark-flavor mixing has been tested in various process
(although in many interesting cases with limited accuracy), finding good agreement with the SM ex-
pectations. The situation is somehow similar to the flavor-conserving electroweak precision observables
after LEP: the SM works very well and genuine one-loop electroweak effects (such as those responsible
for FCNC processes) have been tested with a typical relative accuracy of about 30%. Similarly to the
case of electroweak observables, also in the quark flavor sector non-standard effects can only appear as
small corrections to the leading SM contribution.

Observing new sources of flavor mixing (i.e. flavor violating couplings not related to quark and
lepton mass matrices) is a natural expectation for any extension of the SM with new degrees of freedom
not far from the TeV scale. While direct searches of new particles at high energies provide a direct in-
formation on the mass spectrum of the possible new degrees of freedom, the indirect information from
low-energy flavor-changing processes translates into unique constraints on their couplings. The present
bounds on possible deviations from the SM in flavor-violating processes already set stringent limits on
the flavor structure of physics beyond the SM, and this provides a key information for model-building.
However, several options are still open, and the quality of this information could be substantially im-
proved with improved studies of selected flavor-violating observables. In these lectures we focus on the
interest of future measurements in the $B$- and $D$-meson systems in this perspective.

The lectures are organized as follows: in the first lecture we briefly recall the main features of
flavor physics within the SM. We also address in general terms the so-called flavor problem, namely
the challenge to any SM extension posed by the success of the SM in flavor physics. In the second
lecture we analyse in some detail the SM predictions for some of the most interesting $B$ and $D$ physics
observables to be measured in the LHC era. In the last lecture we analyse flavor physics in various
realistic beyond-the-SM scenarios, discussing how they can be tested by future experiments.

These notes have a sizable overlap with a similar set of lectures I present a few years ago [1]. In-
dependent set of lectures on the same subject can be found in Ref. [2], while more detailed presentations
can be found in the review articles in Ref. [3–6].

1 Flavor physics within the SM and the flavor problem

1.1 The flavor sector of the SM

The Standard Model (SM) Lagrangian can be divided into two main parts, the gauge and the Higgs (or
symmetry breaking) sector. The gauge sector is extremely simple and highly symmetric: it is completely
specified by the local symmetry $G_{\text{local}}^{\text{SM}} = SU(3)_C \times SU(2)_L \times U(1)_Y$ and by the fermion content,

$$
\mathcal{L}_{\text{gauge}}^{\text{SM}} = \sum_{i=1\ldots3} \sum_{\psi=Q^i_L\ldots E^i_R} \bar{\psi} i \not{D} \psi - \frac{1}{4} \sum_{a=1\ldots8} G^{a \mu \nu} G^{a \mu \nu} - \frac{1}{4} \sum_{a=1\ldots3} W^{a \mu \nu} W^{a \mu \nu} - \frac{1}{4} B_{\mu \nu} B_{\mu \nu}.
$$

(1)
The fermion content consist of five fields with different quantum numbers under the gauge group,\(^1\)
\[
Q^i_L(3, 2)_{+1/6}, \quad U^i_R(3, 1)_{+2/3}, \quad D^i_R(3, 1)_{-1/3}, \quad L^i_L(1, 2)_{-1/2}, \quad E^i_R(1, 1)_{-1},
\]
each of them appearing in three different replica or flavors \((i = 1, 2, 3)\).

This structure give rise to a large global flavor symmetry of \(L^{\text{gauge}}_{\text{SM}}\). Both the local and the global symmetries of \(L^{\text{gauge}}_{\text{SM}}\) are broken with the introduction of a \(SU(2)_L\) scalar doublet \(\phi\), or the Higgs field. The local symmetry is spontaneously broken by the vacuum expectation value of the Higgs field, \(\langle \phi \rangle = v = (2\sqrt{2}G_F)^{-1/2} \approx 174 \text{ GeV}\), while the global flavor symmetry is explicitly broken by the Yukawa interaction of \(\phi\) with the fermion fields:
\[
-\mathcal{L}^{\text{Yukawa}} = Y_d^{ij} \bar{Q}^i_L \phi D^j_R + Y_u^{ij} \bar{Q}^i_L \tilde{\phi} U^j_R + Y_e^{ij} \bar{E}^i_L \phi E^j_R + \text{h.c.} \quad (\tilde{\phi} = i\tau_2 \phi^\dagger).
\]
The large global flavor symmetry of \(L^{\text{gauge}}_{\text{SM}}\), corresponding to the independent unitary rotations in flavor space of the five fermion fields in Eq. (2), is a \(U(3)^5\) group. This can be decomposed as follows:
\[
\mathcal{G}_{\text{flavor}} = U(3)^5 \times G_q \times G_\ell,
\]
where
\[
G_q = SU(3)_{Q_L} \times SU(3)_{U_R} \times SU(3)_{D_R}, \quad G_\ell = SU(3)_{L_L} \otimes SU(3)_{E_R}.
\]
Three of the five \(U(1)\) subgroups can be identified with the total baryon and lepton number, which are not broken by \(\mathcal{L}_{\text{Yukawa}}\), and the weak hypercharge, which is gauged and broken only spontaneously by \(\langle \phi \rangle \neq 0\). The subgroups controlling flavor-changing dynamics and flavor non-universality are the non-Abelian groups \(G_q\) and \(G_\ell\), which are explicitly broken by \(Y_{d,u,e}\) not being proportional to the identity matrix.

The diagonalization of each Yukawa coupling requires, in general, two independent unitary matrices, \(V_L Y V_R^\dagger = \text{diag}(y_1, y_2, y_3)\). In the lepton sector the invariance of \(\mathcal{L}^{\text{Yukawa}}_{\text{SM}}\), under \(G_\ell\) allows us to freely choose the two matrices necessary to diagonalize \(Y_\ell\) without breaking gauge invariance, or without observable consequences. This is not the case in the quark sector, where we can freely choose only three of the four unitary matrices necessary to diagonalize both \(Y_d\) and \(Y_u\). Choosing the basis where \(Y_d\) is diagonal (and eliminating the right-handed diagonalization matrix of \(Y_u\)) we can write
\[
Y_d = \lambda_d, \quad Y_u = V^\dagger \lambda_u,
\]
where
\[
\lambda_d = \text{diag}(y_d, y_s, y_b), \quad \lambda_u = \text{diag}(y_u, y_c, y_t), \quad y_q = \frac{m_q}{v}.
\]
Alternatively we could choose a gauge-invariant basis where \(Y_d = V \lambda_d\) and \(Y_u = \lambda_u\). Since the flavor symmetry do not allow the diagonalization from the left of both \(Y_d\) and \(Y_u\), in both cases we are left with a non-trivial unitary mixing matrix, \(V\), which is nothing but the Cabibbo-Kobayashi-Maskawa (CKM) mixing matrix [7, 8].

A generic unitary \(3 \times 3 [N \times N]\) complex unitary matrix depends on three \([N(N - 1)/2]\) real rotational angles and six \([N(N + 1)/2]\) complex phases. Having chosen a quark basis where the \(Y_d\) and \(Y_u\) have the form in (6) leaves us with a residual invariance under the flavor group which allows us to eliminate five of the six complex phases in \(V\) (the relative phases of the various quark fields). As a result, the physical parameters in \(V\) are four: three real angles and one complex CP-violating phase. The full set of parameters controlling the breaking of the quark flavor symmetry in the SM is composed by the six quark masses in \(\lambda_{u,d}\) and the four parameters of \(V\).

\(^1\)The notation used to indicate each field is \(\psi(A, B)_Y\), where \(A\) and \(B\) denote the representation under the \(SU(3)_C\) and \(SU(2)_L\) groups, respectively, and \(Y\) is the \(U(1)_Y\) charge.
For practical purposes it is often convenient to work in the mass eigenstate basis of both up- and down-type quarks. This can be achieved rotating independently the up and down components of the quark doublet $Q_L$, or moving the CKM matrix from the Yukawa sector to the charged weak current in $\mathcal{L}_{\text{SM}}^{\text{gauge}}$:

$$J_W^{\mu}|_{\text{quarks}} = \bar{u}_L \gamma^\mu d_L^{u,d\text{--basis}} \rightarrow \bar{u}_L V_{ij} \gamma^\mu d_L^i .$$

However, it must be stressed that $V$ originates from the Yukawa sector (in particular by the misalignment of $Y_u$ and $Y_d$ in the $SU(3)_{Q_L}$ subgroup of $G_q$): in absence of Yukawa couplings we can always set $V_{ij} = \delta_{ij}$.

To summarize, quark flavor physics within the SM is characterized by a large flavor symmetry, $G_q$, defined by the gauge sector, whose only breaking sources are the two Yukawa couplings $Y_d$ and $Y_u$. The CKM matrix arises from the mis-alignment of $Y_u$ and $Y_d$ in flavor space.

### 1.2 Some properties of the CKM matrix

The standard parametrization of the CKM matrix [9] in terms of three rotational angles ($\theta_{ij}$) and one complex phase ($\delta$) is

$$V = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} = \begin{pmatrix} c_{12} & s_{12}c_{13} & s_{12}e^{i\delta} \\ -s_{12} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta} & c_{12}s_{23} - s_{12}s_{23}s_{13}e^{i\delta} \\ s_{23} & c_{23}s_{13} - c_{23}s_{13}e^{i\delta} & c_{23}s_{13} - c_{23}s_{13}e^{i\delta} \end{pmatrix},$$

where $c_{ij} = \cos \theta_{ij}$ and $s_{ij} = \sin \theta_{ij}$ ($i,j = 1,2,3$).

The off-diagonal elements of the CKM matrix show a strongly hierarchical pattern: $|V_{us}|$ and $|V_{cd}|$ are close to 0.22, the elements $|V_{cb}|$ and $|V_{ts}|$ are of order $4 \times 10^{-2}$ whereas $|V_{ub}|$ and $|V_{td}|$ are of order $5 \times 10^{-3}$. The Wolfenstein parametrization, namely the expansion of the CKM matrix elements in powers of the small parameter $\lambda \doteq |V_{us}| \approx 0.22$, is a convenient way to exhibit this hierarchy in a more explicit way [10]:

$$V = \begin{pmatrix} 1 - \frac{\lambda^2}{2} & \lambda & A\lambda^3 (q - i\eta) \\ -\lambda & 1 - \frac{\lambda^2}{2} & A\lambda^2 \\ A\lambda^3(1 - q - i\eta) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}(\lambda^4),$$

where $A$, $\varrho$, and $\eta$ are free parameters of order 1. Because of the smallness of $\lambda$ and the fact that for each element the expansion parameter is actually $\lambda^2$, this is a rapidly converging expansion.

The Wolfenstein parametrization is certainly more transparent than the standard parametrization. However, if one requires sufficient level of accuracy, the terms of $\mathcal{O}(\lambda^4)$ and $\mathcal{O}(\lambda^5)$ have to be included in phenomenological applications. This can be achieved in many different ways, according to the convention adopted. The simplest (and nowadays commonly adopted) choice is obtained defining the parameters $\{\lambda, A, \varrho, \eta\}$ in terms of the angles of the exact parametrization in Eq. (9) as follows:

$$\lambda \doteq s_{12}, \quad A\lambda^2 \doteq s_{23}, \quad A\lambda^3(q - i\eta) \doteq s_{13}e^{-i\delta}.$$ (11)

The change of variables $\{s_{ij}, \delta\} \rightarrow \{\lambda, A, \varrho, \eta\}$ in Eq. (9) leads to an exact parametrization of the CKM matrix in terms of the Wolfenstein parameters. Expanding this expression up to $\mathcal{O}(\lambda^5)$ leads to

$$\begin{pmatrix} 1 - \frac{1}{2}A^2 + \frac{1}{8}A^4 & \lambda + \mathcal{O}(\lambda^7) & A\lambda^3(q - i\eta) \\ -\lambda + \frac{1}{2}A^2\lambda^3[1 - 2(\varrho + i\eta)] & 1 - \frac{1}{2}A^2 - \frac{1}{8}A^4(1 + 4A^2) & A\lambda^2 + \mathcal{O}(\lambda^6) \\ A\lambda^3(1 - \varrho - i\eta) & -A\lambda^2 + \frac{1}{2}A^2\lambda^4[1 - 2(\varrho + i\eta)] & 1 - \frac{1}{2}A^2\lambda^4 \end{pmatrix}$$
The CKM unitarity triangle

where

\[ \bar{\rho} = \rho \left(1 - \frac{\lambda^2}{2}\right) + \mathcal{O}(\lambda^4), \quad \bar{\eta} = \eta \left(1 - \frac{\lambda^2}{2}\right) + \mathcal{O}(\lambda^4). \] (13)

The advantage of this generalization of the Wolfenstein parametrization is the absence of relevant corrections to \( V_{us}, V_{cd}, V_{ab} \) and \( V_{cb} \), and a simple change in \( V_{td} \), which facilitate the implementation of experimental constraints.

The unitarity of the CKM matrix implies the following relations between its elements:

\[
\sum_{k=1\ldots3} V_{ik}^* V_{ki} = 1, \quad \sum_{k=1\ldots3} V_{ik}^* V_{kj} \neq i \text{.} \tag{14}
\]

These relations are a distinctive feature of the SM, where the CKM matrix is the only source of quark flavor mixing. Their experimental verification is therefore a useful tool to set bounds, or possibly reveal, new sources of flavor symmetry breaking. Among the relations of type II, the one obtained for \( i = 1 \) and \( j = 3 \), namely

\[ V_{ud} V_{ub}^* + V_{cd} V_{cb}^* + V_{td} V_{tb}^* = 0 \] (15)

or

\[
\frac{V_{ud} V_{ub}^*}{V_{cd} V_{cb}^*} + \frac{V_{id} V_{ib}^*}{V_{cd} V_{cb}^*} + 1 = 0 \quad \leftrightarrow \quad [\bar{\rho} + i\bar{\eta}] + [(1 - \bar{\rho}) - i\bar{\eta}] + 1 = 0 ,
\]

is particularly interesting since it involves the sum of three terms all of the same order in \( \lambda \) and is usually represented as a unitarity triangle in the complex plane, as shown in Fig. 1. It is worth to stress that Eq. (15) is invariant under any phase transformation of the quark fields. Under such transformations the triangle in Fig. 1 is rotated in the complex plane, but its angles and the sides remain unchanged. Both angles and sides of the unitary triangle are indeed observable quantities which can be measured in suitable experiments.

1.3 Present status of CKM fits

The values of \( |V_{us}| \) and \( |V_{cb}| \), or \( \lambda \) and \( A \) in the parametrization (12), are determined with good accuracy from \( K \to \pi \ell \nu \) and \( B \to X_c \ell \nu \) decays, respectively. According to the recent analysis of the UTfit collaboration [13] their numerical values are

\[ \lambda = 0.2259 \pm 0.0006, \quad A = 0.824 \pm 0.013 \] (16)

Using these results, all the other constraints on the elements of the CKM matrix can be expressed as constraints on \( \bar{\rho} \) and \( \bar{\eta} \) (or constraints on the CKM unitarity triangle in Fig. 1). The list of the most sensitive observables used to determine \( \bar{\rho} \) and \( \bar{\eta} \) in the SM includes:

- The rates of inclusive and exclusive charmless semileptonic \( B \) decays, that depend on \( |V_{ub}| \) and provide a constraint on \( \bar{\rho}^2 + \bar{\eta}^2 \).
The time-dependent CP asymmetry in $B \to \psi K_S$ decays ($A_{\psi K_S}^{\text{CP}}$), that depends on the phase of the $B_d$-$\bar{B}_d$ mixing amplitude relative to the decay amplitude (see Section 2.2). Within the SM this translates into a constraint on $\sin 2\beta$.

- The rates of various $B \to D K$ decays constraining the angle $\gamma$ (see Section 2.3).
- The rates of various $B \to \pi\pi, \rho\pi, \rho\rho$ decays constraining the combination $\alpha = \pi - \beta - \gamma$.
- The ratio between the mass splittings in the neutral $B$ and $B_s$ systems, that depends on $|V_{td}/V_{ts}|^2 \propto [(1 - \bar{\rho})^2 + \bar{\eta}^2]$.
- The indirect CP violating parameter of the kaon system ($\epsilon_K$), that determines and hyperbola in the $\bar{\rho}$ and $\bar{\eta}$ plane (see Ref. [3] for more details).

The resulting constraints, as implemented by the CKMfitter collaboration, are shown in Fig. 2. As can be seen, they are all consistent with a unique value of $\bar{\rho}$ and $\bar{\eta}$ (the results obtained at present by the two most representative fitter groups, the CKMfitter and the UTfit collaboration, are in good agreement). The numerical values for the best fit values of $\bar{\rho}$ and $\bar{\eta}$ quoted in Ref. [13] are

$$\bar{\rho} = 0.142 \pm 0.022 , \quad \bar{\eta} = 0.352 \pm 0.016 .$$  (17)
1.4 The SM as an effective theory

As anticipated in the introduction, despite the impressive phenomenological success of the SM in flavor and electroweak physics, there are various convincing arguments which motivate us to consider this model only as the low-energy limit of a more complete theory.

Assuming that the new degrees of freedom which complete the theory are heavier than the SM particles, we can integrate them out and describe physics beyond the SM in full generality by means of an effective theory approach. The SM Lagrangian becomes the renormalizable part of a more general local Lagrangian which includes an infinite tower of operators with dimension \( d > 4 \), constructed in terms of SM fields and suppressed by inverse powers of an effective scale \( \Lambda \). These operators are the residual effect of having integrated out the new heavy degrees of freedom, whose mass scale is parametrized by the effective scale \( \Lambda > m_W \).

As we will discuss in more detail in Section 2.1.1, integrating out heavy degrees of freedom is a procedure often adopted also within the SM: it allows us to simplify the evaluation of amplitudes which involve different energy scales. This approach is indeed a generalization of the Fermi theory of weak interactions, where the dimension-six four-fermion operators describing weak decays are the results of having integrated out the \( W \) field. The only difference when applying this procedure to physics beyond the SM is that in this case, as also in the original work by Fermi, we don’t know the nature of the degrees of freedom we are integrating out. This imply we are not able to determine a priori the values of the effective couplings of the higher-dimensional operators. The advantage of this approach is that it allows us to analyse all realistic extensions of the SM in terms of a limited number of parameters (the coefficients of the higher-dimensional operators). The drawback is the impossibility to establish correlations of New Physics (NP) effects at low and high energies.

Assuming for simplicity that there is a single elementary Higgs field, responsible for the \( SU(2)_L \times U(1)_Y \rightarrow U(1)_Q \) spontaneous breaking, the Lagrangian of the SM considered as an effective theory can be written as follows

\[
L_{\text{eff}} = L_{\text{gauge}}^{\text{SM}} + L_{\text{Higgs}}^{\text{SM}} + L_{\text{Yukawa}}^{\text{SM}} + \Delta L_{d>4} ,
\]

where \( \Delta L_{d>4} \) denotes the series of higher-dimensional operators invariant under the SM gauge group:

\[
\Delta L_{d>4} = \sum_{d>4} \sum_{n=1}^{N_d} \frac{c^{(d)}_{n}}{\Lambda^{d-4}} O^{(d)}_{n} (\text{SM fields}).
\]

If NP appears at the TeV scale, as we expect from the stabilization of the mechanism of electroweak symmetry breaking, the scale \( \Lambda \) cannot exceed a few TeV. Moreover, if the underlying theory is natural (no fine-tuning in the coupling constants), we expect \( c^{(d)}_{n} = O(1) \) for all the operators which are not forbidden (or suppressed) by symmetry arguments. The observation that this expectation is not fulfilled by several dimension-six operators contributing to flavor-changing processes is often denoted as the flavor problem.

If the SM Lagrangian were invariant under some flavor symmetry, this problem could easily be circumvented. For instance in the case of barion- or lepton-number violating processes, which are exact symmetries of the SM Lagrangian, we can avoid the tight experimental bounds promoting \( B \) and \( L \) to be exact symmetries of the new dynamics at the TeV scale. The peculiar aspects of flavor physics is that there is no exact flavor symmetry in the low-energy theory. In this case it is not sufficient to invoke a flavor symmetry for the underlying dynamics. We also need to specify how this symmetry is broken in order to describe the observed low-energy spectrum and, at the same time, be in agreement with the precise experimental tests of flavor-changing processes.

1.4.1 Bounds on the scale of New Physics from \( \Delta F = 2 \) processes

The best way to quantify the flavor problem is obtained by looking at consistency of the tree- and loop-mediated constraints on the CKM matrix discussed in Section 1.3.
In first approximation we can assume that NP effects are negligible in processes which are dominated by tree-level amplitudes. Following this assumption, the values of $|V_{uS}|$, $|V_{cb}|$, and $|V_{ub}|$, as well as the constraints on $\alpha$ and $\gamma$ are essentially NP free. As can be seen in Fig. 2, this implies we can determine completely the CKM matrix assuming generic NP effects in loop-mediated amplitudes. We can then use the measurements of observables which are loop-mediated within the SM to bound the couplings of effective NP operators in Eq. (19) which contribute to these observables at the tree level.

The loop-mediated constraints shown in Fig. 2 are those from the mixing of $B_d$, $B_s$, and $K^0$ with the corresponding anti-particles (generically denoted as $\Delta F = 2$ amplitudes). Within the SM, these amplitudes are generated by box amplitudes of the type in Fig. 3 (and similarly for $B_s$, and $K^0$) and are affected by small hadronic uncertainties (with the exception of $\Delta m_K$). We will come back to the evaluation of these amplitudes in more detail in Section 2.2. For the moment it is sufficient to notice that the leading contribution is obtained with the top-quark running inside the loop, giving rise to the highly suppressed result

$$M_{\Delta F=2}^{SM} \approx \frac{G_F^2 m_t^2}{16\pi^2} V_{td}^* V_{tb} \langle M | (d_L^\gamma \gamma^\mu d_L^\nu)^2 | M \rangle \times \frac{m_t^2}{m_W^2} | M = K^0, B_d, B_s \rangle,$$

where $F$ is a loop function of order one ($i, j$ denote the flavor indexes of the meson valence quarks).

Magnitude and phase of all these mixing amplitudes have been determined with good accuracy from experiments with the exception of the CP-violating phase in $B_s-B_s$ mixing. As shown in Fig. 2, in all cases where the experimental information is precise, the magnitude of the new-physics amplitude cannot exceed, in size, the SM contribution.

To translate this information into bounds on the scale of new physics, let’s consider the following set of $\Delta F = 2$ dimensions-six operators

$$O_{\Delta F=2}^{ij} = (Q_L^i \gamma^\mu Q_L^j)^2, \quad Q_L^i = \left( \begin{array}{c} u_L^i \\ d_L^i \end{array} \right),$$

where $i, j$ are flavor indexes in the basis defined by Eq. (6). These operators contribute at the tree-level to the meson-antimeson mixing amplitudes. Denoting $c_{ij}$ the couplings of the non-standard operators in (21), the condition $|M_{\Delta F=2}^{NP}| < |M_{\Delta F=2}^{SM}|$ implies

$$\Lambda < \frac{3.4 \text{ TeV}}{|V_{td}^* V_{tb}|/|c_{ij}|^{1/2}} < \begin{cases} 9 \times 10^3 \text{ TeV} \times |c_{21}|^{1/2} & \text{from } K^0 - \bar{K}^0 \\ 4 \times 10^2 \text{ TeV} \times |c_{31}|^{1/2} & \text{from } B_d - \bar{B}_d \\ 7 \times 10^1 \text{ TeV} \times |c_{32}|^{1/2} & \text{from } B_s - \bar{B}_s \end{cases}$$

A more refined analysis, with complete statistical treatment and separate bounds for the real and the imaginary parts of the various amplitudes, considering also operators with different Dirac structure, is reported in Table 1.\textsuperscript{2} The main messages of these bounds are the following:

- New physics models with a generic flavor structure ($c_{ij}$ of order 1) at the TeV scale are ruled out.
- If we want to keep $\Lambda$ in the TeV range, physics beyond the SM must have a highly non-generic flavor structure.

\textsuperscript{2}Table 1 updates the corresponding table of Ref. [5] taking into account the recent measurements in the $B_s$ system.
Table 1: Bounds on representative dimension-six $\Delta F = 2$ operators, assuming an effective coupling $c_{NP}/\Lambda^2$. The bounds are quoted on $\Lambda$, setting $|c_{NP}| = 1$, or on $c_{NP}$, setting $\Lambda = 1$ TeV. The right column denotes the main observables used to derive these bounds (see next section for more details).

<table>
<thead>
<tr>
<th>Operator</th>
<th>Bounds on $\Lambda$ in TeV ($c_{NP} = 1$)</th>
<th>Bounds on $c_{NP}$ ($\Lambda = 1$ TeV)</th>
<th>Observables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Re</td>
<td>Im</td>
<td>Re</td>
</tr>
<tr>
<td>($\bar{s}_L \gamma^\mu d_L$)$^2$</td>
<td>$9.8 \times 10^4$</td>
<td>$1.6 \times 10^4$</td>
<td>$9.0 \times 10^{-7}$</td>
</tr>
<tr>
<td>($\bar{s}_L d_L$)(\bar{s}_L d_R)</td>
<td>$1.8 \times 10^4$</td>
<td>$3.2 \times 10^5$</td>
<td>$6.9 \times 10^{-9}$</td>
</tr>
<tr>
<td>($\bar{c}_L \gamma^\mu u_L$)$^2$</td>
<td>$1.2 \times 10^3$</td>
<td>$2.9 \times 10^3$</td>
<td>$5.6 \times 10^{-7}$</td>
</tr>
<tr>
<td>($\bar{c}_L u_L$)(\bar{c}_L u_R)</td>
<td>$6.2 \times 10^3$</td>
<td>$1.5 \times 10^4$</td>
<td>$5.7 \times 10^{-8}$</td>
</tr>
<tr>
<td>($b_L \gamma^\mu d_L$)$^2$</td>
<td>$6.0 \times 10^2$</td>
<td>$9.3 \times 10^2$</td>
<td>$2.3 \times 10^{-6}$</td>
</tr>
<tr>
<td>($\bar{b}_L d_L$)(\bar{b}_L d_R)</td>
<td>$2.5 \times 10^3$</td>
<td>$3.6 \times 10^3$</td>
<td>$3.9 \times 10^{-7}$</td>
</tr>
<tr>
<td>($b_L \gamma^\mu s_L$)$^2$</td>
<td>$1.4 \times 10^2$</td>
<td>$2.5 \times 10^2$</td>
<td>$5.0 \times 10^{-5}$</td>
</tr>
<tr>
<td>($\bar{b}_L s_L$)(\bar{b}_L s_R)</td>
<td>$4.8 \times 10^2$</td>
<td>$8.3 \times 10^2$</td>
<td>$8.8 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

In the specific case of the $\Delta F = 2$ operators in (21), in order to keep $\Lambda$ in the TeV range, we must find a symmetry argument such that $|c_{ij}| \lesssim |V_{3i}^* V_{3j}|^2$.

The strong constraining power of $\Delta F = 2$ observables is a consequence of their strong suppression within the SM. They are suppressed not only by the typical $1/(4\pi)^2$ factor of loop amplitudes, but also by the GIM mechanism [14] and by the hierarchy of the CKM matrix ($|V_{3i}| \ll 1$, for $i \neq 3$). Reproducing a similar structure beyond the SM is a highly non-trivial task. As we will discuss in the last lecture, only in a few cases this can be implemented in a natural way.

To conclude, we stress that the good agreement of SM and experiments for $B_d$ and $K^0$ mixing does not imply that further studies of flavor physics are not interesting. On the one hand, even for $|c_{ij}| \approx |V_{3i}^* V_{3j}|$, which can be considered the most pessimistic case, as we will discuss in Section 3.1, we are presently constraining physics at the TeV scale. Therefore improving these bounds, if possible, would be extremely valuable. One the other hand, as we will discuss in the next lecture, there are various interesting observables which have not been deeply investigated yet, whose study could reveal additional key features about the flavor structure of physics beyond the SM.

## 2 Phenomenology of $B$ and $D$ decays

As we have seen in the previous lecture, the exploration of the mechanism of quark-flavor mixing is entered in a new era. The precise measurements of mixing-induced CP violation and tree-level allowed semileptonic transition have provided an important consistency check of the SM, and a precise determination of the Cabibbo-Kobayashi-Maskawa matrix. The next goal is to understand if there is still room for new physics or, more precisely, if there is still room for new sources of flavor symmetry breaking close to the electroweak scale. From this perspective, the meson-antimeson mixing amplitudes, CP-violating observables, and the rates to few rare $B$ decays mediated by flavor-changing neutral-current (FCNC) represent a fundamental tool.

Beside the experimental sensitivity, the conditions which allow us to perform significant NP searches in rare decays can be summarized as follows: i) decay amplitude dominated by electroweak dynamics, and thus enhanced sensitivity to non-standard contributions; ii) small theoretical error within the SM, or good control of both perturbative and non-perturbative corrections.

In this lecture we first we introduce the main theoretical tools that allow us to evaluate at which level these two conditions are satisfied in a given observable. We then apply these tool to analyse in more detail a few selected observables: i) the determination of the CP-violating phase of the $B_s$ mixing
amplitude; ii) the determination of the CKM phase $\gamma$ from charged $B \to DK$ decays; iii) the rare decays $B_{s,d} \to \ell^+\ell^-$; iv) CP violation in $D$ decays. The selection is far from being exhaustive (for a more complete analysis we refer to the reviews in Ref. [3, 4]), but it should serve as an illustration of the interesting potential of $B$ and $D$ physics at hadron colliders.

2.1 Theoretical tools

2.1.1 Low-energy effective Lagrangians

The decays of $B$ mesons are processes which involve at least two different energy scales: the electroweak scale, characterized by the $W$ boson mass, which determines the flavor-changing transition at the quark level, and the scale of strong interactions $\Lambda_{\text{QCD}}$, related to the hadron formation. The presence of these two widely separated scales makes the calculation of the decay amplitudes starting from the full SM Lagrangian quite complicated: large logarithms of the type $\log(m_W/\Lambda_{\text{QCD}})$ may appear, leading to a breakdown of ordinary perturbation theory.

This problem can be substantially simplified by integrating out the heavy SM fields ($W$ and $Z$ bosons, as well as the top quark) at the electroweak scale, and constructing an appropriate low-energy effective theory where only the light SM fields appear. The weak effective Lagrangians thus obtained contains local operators of dimension six (and higher), written in terms of light SM fermions, photon and gluon fields, suppressed by inverse powers of the $W$ mass.

To be concrete, let’s consider the example of charged-current semileptonic weak interactions. The basic building block in the full SM Lagrangian is

$$L_{W}^{\text{full SM}} = \frac{g}{\sqrt{2}} J_{W}^{\mu}(x) W_{\mu}^{+}(x) + \text{h.c.}, \quad \text{(23)}$$

where

$$J_{W}^{\mu}(x) = V_{ij} \bar{u}_{L}^{i}(x) \gamma^{\mu} d_{L}^{j}(x) + \bar{e}_{L}^{i}(x) \gamma^{\mu} \nu_{L}^{j}(x) \quad \text{(24)}$$

is the weak charged current already introduced in Eq. (8). Integrating out the $W$ field at the tree level we contract two vertexes of this type generating the non-local transition amplitude

$$iT = -i g_{\mu}^{2} \int d^{4}x D_{\mu\nu}(x, m_{W}) T \left[ J_{W}^{\mu}(x), J_{W}^{\nu}(0) \right], \quad \text{(25)}$$

which involves only light fields. Here $D_{\mu\nu}(x, m_{W})$ is the $W$ propagator in coordinate space: expanding it in inverse powers of $m_{W}$,

$$D_{\mu\nu}(x, m_{W}) = \int \frac{d^{4}q}{(2\pi)^{4}} e^{-iq \cdot x} - i g_{\mu\nu} + \mathcal{O}(q_{\mu}, q_{\nu}) = \delta(x) \frac{i g_{\mu\nu}}{m_{W}^{2}} + \ldots, \quad \text{(26)}$$

the leading contribution to $T$ can be interpreted as the tree-level contribution of the following effective local Lagrangian

$$L_{\text{eff}}^{(0)} = -\frac{4G_{F}}{\sqrt{2}} g_{\mu\nu} J_{W}^{\mu}(x) J_{W}^{\nu}(x), \quad \text{(27)}$$

where $G_{F}/\sqrt{2} = g^{2}/(8m_{W}^{2})$ is the Fermi coupling. If we select in the product of the two currents one quark and one leptonic current,

$$L_{\text{eff}}^{\text{semi-lept}} = -\frac{4G_{F}}{\sqrt{2}} V_{ij} \bar{u}_{L}^{i}(x) \gamma^{\mu} d_{L}^{j}(x) \bar{\nu}_{L}(x) \gamma_{\mu} \nu_{L}(x) + \text{h.c.}, \quad \text{(28)}$$

we obtain an effective Lagrangian which provides an excellent description of semileptonic weak decays. The neglected terms in the expansion (26) correspond to corrections of $\mathcal{O}(m_{B}^{2}/m_{W}^{2})$ to the decay amplitudes. In principle, these corrections could be taken into account by adding appropriate dimension-eight operators in the effective Lagrangian. However, in most cases they are safely negligible.

G. ISIDORI
The case of charged semileptonic decays is particularly simple since we can ignore QCD effects: the operator (28) is not renormalized by strong interactions. The situation is slightly more complicated in the case of non-leptonic or flavor-changing neutral-current processes, where QCD corrections and higher-order weak interactions cannot be neglected, but the basic strategy is the same. First of all we need to identify a complete basis of local operators, that includes also those generated beyond the tree level. In general, given a fixed order in the $1/m_W^2$ expansion of the amplitudes, we need to consider all operators of corresponding dimension (e.g. dimension six at the first order in the $1/m_W^2$ expansion) compatible with the symmetries of the system. Then we must introduce an artificial scale in the problem, the renormalization scale $\mu$, which is needed to regularize QCD (or QED) corrections in the effective theory.

The effective Lagrangian for generic $\Delta F = 1$ processes assumes the form

$$\mathcal{L}_{\Delta F=1} = -4 \frac{G_F}{\sqrt{2}} \sum_i C_i(\mu) Q_i$$

where the sum runs over the complete basis of operators. As explicitly indicated, the effective couplings $C_i(\mu)$ (known as Wilson coefficients) depend, in general, on the renormalization scale. The dependence from this scale cancels when evaluating the matrix elements of the effective Lagrangian for physical processes, that we can generically indicate as

$$\mathcal{M}(i \to f) = -4 \frac{G_F}{\sqrt{2}} \sum_i C_i(\mu) \langle f | Q_i(\mu) | i \rangle.$$  

The independence of $\mathcal{M}$ from $\mu$ holds for any initial and final state, including partonic states at high energies. This implies that the $C_i(\mu)$ obey a series of renormalization group equations (RGE), whose structure is completely determined by the anomalous dimensions of the effective operators. These equations can be solved using standard RG techniques, allowing the resummation of all large logs of the type $\alpha_s(\mu)^{n+1} \log(m_W/\mu)^n$ to all orders in $n$ (working at order $m + 1$ in perturbation theory). The scale $\mu$ acts as a separator of short- and long-distance virtual corrections: short-distance effects are included in the $C_i(\mu)$, whereas long-distance effects are left as explicit degrees of freedom in the effective theory.\footnote{This statement would be correct if the theory were regularized using a dimensional cut-off. It is not fully correct if $\mu$ is the scale appearing in the (often adopted) dimensional-regularization + minimal-subtraction (MS) renormalization scheme.}

In practice, the problem reduces to the following three well-defined and independent steps:

1. the evaluation of the initial conditions of the $C_i(\mu)$ at the electroweak scale ($\mu \approx m_W$);
2. the evaluation of the anomalous dimension of the effective operators, and the corresponding RGE evolution of the $C_i(\mu)$ from the electroweak scale down to the energy scale of the physical process ($\mu \approx m_B$);
3. the evaluation of the matrix elements of the effective Lagrangian for the physical hadronic processes (which involve energy scales from $m_B$ down to $\Lambda_{QCD}$).

The first step is the one where physics beyond the SM may appear: if we assume NP is heavy, it may modify the initial conditions of the Wilson coefficients at the high scale, while it cannot affect the following two steps. While the RGE evolution and the hadronic matrix elements are not directly related to NP, they may influence the sensitivity to NP of physical observables. In particular, the evaluation of hadronic matrix elements is potentially affected by non-perturbative QCD effects: these are often a large source of theoretical uncertainty which can obscure NP effects. RGE effects do not induce sizable uncertainties since they can be fully handled within perturbative QCD; however, the sizable logs generated by the RGE running may dilute the interesting short-distance information encoded in the value of the Wilson coefficients at the high scale. As we will discuss in the following, only in specific observables these two effects are small and under good theoretical control.
A deeper discussion about the construction of low-energy effective Lagrangians, with a detailed discussion of the first two steps mentioned above, can be found in Ref. [15].

2.1.1.1 Effective operators for rare processes

Let’s give a closer look to processes where the underlying parton process is $b \to s + \bar{q} q$. In this case the relevant effective Lagrangian can be written as

$$\mathcal{L}_{b \to s}^{\text{non-lept}} = -4 \frac{G_F}{\sqrt{2}} \left( \sum_{q=u,c} \lambda_q^s \sum_{i=1}^{10} C_i(\mu) Q_i^a(\mu) - \lambda_q^s \sum_{i=3}^{12} C_i(\mu) Q_i(\mu) \right),$$

where $\lambda_q^s = V_{qs}^* V_{qs}$, and the operator basis is

$$Q_1^a = \frac{2}{3} b_L^\alpha \gamma^\mu q_L^i \bar{q}_L^\gamma \gamma_\mu s_L^i,$$

$$Q_2^a = \frac{2}{3} b_L^\alpha \gamma^\mu q_L^i \bar{q}_L^\gamma \gamma_\mu s_L^i,$$

$$Q_3^a = \frac{2}{3} b_L^\alpha \gamma^\mu q_L^i \bar{q}_L^\gamma \gamma_\mu s_L^i,$$

$$Q_4^a = \frac{2}{3} b_L^\alpha \gamma^\mu q_L^i \bar{q}_L^\gamma \gamma_\mu s_L^i,$$

$$Q_5^a = \frac{2}{3} b_L^\alpha \gamma^\mu q_L^i \bar{q}_L^\gamma \gamma_\mu s_L^i,$$

$$Q_6^a = \frac{2}{3} b_L^\alpha \gamma^\mu q_L^i \bar{q}_L^\gamma \gamma_\mu s_L^i,$$

$$Q_7^a = \frac{2}{3} b_L^\alpha \gamma^\mu q_L^i \bar{q}_L^\gamma \gamma_\mu s_L^i,$$

$$Q_8^a = \frac{2}{3} b_L^\alpha \gamma^\mu q_L^i \bar{q}_L^\gamma \gamma_\mu s_L^i,$$

$$Q_9^a = \frac{2}{3} b_L^\alpha \gamma^\mu q_L^i \bar{q}_L^\gamma \gamma_\mu s_L^i,$$

$$Q_{10}^a = \frac{2}{3} b_L^\alpha \gamma^\mu q_L^i \bar{q}_L^\gamma \gamma_\mu s_L^i,$$

with $\{\alpha, \beta\}$ and $e_q$ denoting color indexes and the electric charge of the quark $q$, respectively.

Out of these operators, only $Q_1^a$ and $Q_4^a$ are generated at the tree-level by the $W$ exchange. Indeed, comparing with the tree-level structure in (27), we find

$$C_{1}^{u,c}(m_W) = 1 + O(\alpha_s, \alpha), \quad C_{2}^{u,c}(m_W) = 0 + O(\alpha_s, \alpha).$$

However, after including RGE effects and running down to $\mu \approx m_b$, both $C_{1}^{u,c}$ and $C_{2}^{u,c}$ become $O(1)$, while $C_{3-6}$ become $O(\alpha_s(m_b))$. In all these cases there is little hope to identify NP effects: the leading initial condition is the tree-level $W$ exchange, which is hardly modified by NP. In principle, the coefficients of the electroweak penguin operators, $Q_{7}-Q_{10}$, are more interesting: their initial conditions are related to electroweak penguin and box diagrams. However, it is hard to distinguish their contribution from those of the other four-quark operators in non-leptonic processes. Moreover, also for $C_{7-10}$ the relative contribution from long-distance physics (running down from $m_W$ to $m_b$) is sizable and dilute the interesting short-distance information.

For $b \to s$ transitions with a photon or a lepton pair in the final state, additional dimension-six operators must be included in the basis,

$$\mathcal{L}_{b \to s}^{\text{rare}} = \mathcal{L}_{b \to s}^{\text{non-lept}} + 4 \frac{G_F}{\sqrt{2}} \lambda_t^s \left( C_{7} Q_{7} + C_{8g} Q_{8g} + C_{9V} Q_{9V} + C_{10A} Q_{10A} \right),$$

where

$$Q_{7} = \frac{e}{16\pi^2} m_b \bar{b}^\alpha \gamma_\mu T^\alpha F_{\mu\nu} s_L^i,$$

$$Q_{8g} = \frac{g_s}{16\pi^2} m_b \bar{b}^\alpha \gamma_\mu T^\alpha A_{\mu} s_L^i,$$

$$Q_{9V} = \frac{1}{2} \bar{b}_L^\alpha \gamma_\mu s_L^i \gamma_\nu l,$$

$$Q_{10A} = \frac{1}{2} \bar{b}_L^\alpha \gamma_\mu s_L^i \gamma_\nu \gamma_5 l,$$

and $G_{\mu\nu}^A$ ($F_{\mu\nu}$) is the gluon (photon) field strength tensor. The initial conditions of these operators are particularly sensitive to NP; within the SM they are generated by one-loop penguin and box diagrams dominated by the top-quark exchange. The most theoretically clean is $C_{10A}$, which do not mix with any of the four-quark operators listed above and which has a vanishing anomalous dimension:

$$C_{10A}^{\text{SM}}(m_W) = \frac{g}{8\pi^2} \frac{x_t}{8} \left( \frac{4 - x_t}{1 - x_t} + \frac{3x_t}{(1 - x_t)^2} \ln x_t \right), \quad x_t = \frac{m_t^2}{m_W^2}. $$

80
NP effects at the TeV scale could easily modify this result, and this deviation would directly show up in low-energy observables sensitive to $C_{10A}$, such as $A_{FB}(B \to K^* \ell^+ \ell^-)$ and $B(B \to \ell^+ \ell^-)$ (see Sections 2.4.1 and 2.4.2). We finally note that while the operators in Eqs. (32) and (35) form a complete basis within the SM, this is not necessarily the case beyond the SM. In particular, within specific scenarios also right-handed current operators (e.g. those obtained from (35) for $q_{L(R)} \to q_{R(L)}$) may appear.

2.1.1.2 Effective operators for meson-antimeson mixing

The $\Delta F = 2$ effective weak Lagrangians are simpler than the $\Delta F = 1$ ones: the SM operator basis includes one operator only. The Lagrangian relevant for $B^0_d - \bar{B}^0_d$ and $B^0_s - \bar{B}^0_s$ mixing is conventionally written as ($q = \{d, s\}$):

$$\mathcal{L}^{\Delta B=2}_{\Delta B=2} = \frac{G_F^2 m_W^2}{4\pi^2} (V_{tb}^* V_{tb})^2 \eta_B(\mu) S_0(x_t) (\bar{b}_L \gamma_\mu q_L \bar{b}_L \gamma_\mu q_L),$$

(37)

where the initial condition of the Wilson coefficient is the loop function $S_0(x_t)$, corresponding to the box diagrams in Fig. 3. The effect of QCD correction is only a multiplicative correction factor, $\eta_B(\mu)$, which can be computed with high accuracy and turns out to be of order one. The explicit expression of the loop function, dominated by the top-quark exchange, is

$$S_0(x_t) = \frac{4x_t - 11x_t^2 + x_t^3}{4(1 - x_t)^2} - \frac{3x_t^2 \ln x_t}{2(1 - x_t)^3}. \quad (38)$$

2.1.2 The gauge-less limit of FCNC amplitudes

An interesting aspect which is common to the electroweak loop functions in Eqs. (36) and (38) is the fact they diverge in the limit $m_t/m_W \to \infty$. This behavior is apparently strange: it contradicts the expectation that contributions of heavy particles at low energy decouple in the limit where their masses increase. The origin of this effect can be understand by noting that the leading contributions to both amplitudes are generated only by the Yukawa interaction. These contributions can be better isolated in the gaugeless limit of the SM, i.e. if we send to zero the gauge couplings. In this limit $m_W \to 0$ and the derivation of the effective Lagrangian discussed in Section 2.1.1 does not make sense. However, the leading contributions to the effective Lagrangians for $\Delta F = 2$ and rare decays are unaffected. Indeed, the leading contributions to these processes are generated by Yukawa interactions of the type in Fig. 4, where the scalar fields are the Goldstone-bosons components of the Higgs field (which are not eaten up by the $W$ in the limit $g \to 0$). Since the top is still heavy, we can integrate it out, obtaining the following result for $\mathcal{L}^{\Delta B=2}_{\Delta B=2}$:

$$\mathcal{L}^{\Delta B=2}_{\Delta B=2}|_{g_t \to 0} = \frac{G_F^2 m_t^2}{16\pi^2} (V_{tb}^* V_{tb})^2 (\bar{b}_L \gamma_\mu q_L)^2 = \frac{[Y_u Y_{u'}^* q_L]^2}{128\pi^2 m_t^2} (\bar{b}_L \gamma_\mu q_L)^2. \quad (39)$$

Taking into account that $S_0(x) \to x/4$ for $x \to \infty$, it is easy to verify that this result is equivalent to the one in Eq. (38) in the large $m_t$ limit. A similar structure holds for the $\Delta F = 1$ amplitude contributing to the axial operator $Q_{10A}$. 
The last expression in Eq. (39), which holds in the limit where we neglect the charm Yukawa coupling, shows that the decoupling of the amplitude with the mass of the top is compensated by four powers of the top Yukawa coupling at the numerator. The divergence for \( m_t \to \infty \) can thus be understood as the divergence of one of the fundamental couplings of the theory. Note also that in the gaugeless limit there is no GIM mechanism: the contributions of the various up-type quarks inside the loops do not cancel each other: they are directly weighted by the corresponding Yukawa couplings, and this is why the top-quark contribution is the dominant one.

This exercise illustrates the key role of the Yukawa coupling in determining the main properties flavor physics within the SM, as advertised in the first lecture. It also illustrates the interplay of flavor and electroweak symmetry breaking in determining the structure of short-distance dominated flavor-changing processes in the SM.

### 2.1.3 Hadronic matrix elements

As anticipated, all non-perturbative effects are confined in the hadronic matrix elements of the operators of the effective Lagrangians. As far as the evaluation of the matrix elements is concerned, we can divide \( B \)-physics observables in three main categories: i) inclusive decays, ii) one-hadron final states, iii) multi-hadron processes.

The heavy-quark expansion [16] form a solid theoretical framework to evaluate the hadronic matrix elements for inclusive processes: inclusive hadronic rates are related to those of free \( b \) quarks, calculable in perturbation theory, by means of a systematic expansion in inverse powers of \( \Lambda_{\text{QCD}}/m_b \). Thanks to quark-hadron duality, the lowest-order terms in this expansion are the pure partonic rates, and for sufficiently inclusive observables higher-order corrections are usually very small. This technique has been very successful in the past in the case of charged-current semileptonic decays, as well as \( B \to X_s \gamma \). However, it has a limited domain of applicability, due to the difficulty of selecting and reconstructing hadronic inclusive states. It cannot be used at hadronic machines, and even at \( B \) factories it cannot be applied to very rare decays.

For processes with a single hadron in the final state, the hadronic effects are often (although not always) confined to the matrix elements of a single quark current. These can be expressed in terms of the meson decay constants

\[
\langle 0 | b \gamma \gamma_5 q | B_q(p) \rangle = i p_{\mu} F_{B_q} ,
\]

or appropriate \( B \to H \) hadronic form factors. Lattice QCD is the best tool to evaluate these non-perturbative quantities from first principles, at least in the kinematical region where the form factors are real (no re-scattering phase allowed). At present not all the form-factors relevant for \( B \)-physics phenomenology are computed on the lattice with good accuracy, but the field is evolving rapidly (see Ref. [17, 18]). To this category belong also the so-called bag-parameters for \( \Delta B = 2 \) mixing, \( B_{d,s} \), defined by

\[
\eta_B(\mu) \langle \hat{B}_{q} \rangle (\hat{b} L \gamma_\mu q L)^2 | B_q \rangle = \frac{2}{3} f_{B_q}^2 m_{B_q}^2 \eta_B(\mu) B_q(\mu) = \frac{2}{3} f_{B_q}^2 m_{B_q}^2 \hat{\eta}_B \hat{B}_q ,
\]

where both \( \hat{B}_q \) and \( \hat{\eta}_B \) are scale-independent quantities (\( \hat{\eta}_B = 0.55 \pm 0.01 \)). For later convenience, we report here some lattice averages for meson decay constants and bag parameters:

\[
F_{B_s} = 227 \pm 8 \text{ GeV}, \quad \hat{B}_s = 1.22 \pm 0.12 ,
\]

\[
F_{B_d} = 189 \pm 8 \text{ GeV}, \quad \frac{\hat{B}_{B_s}}{\hat{B}_{B_d}} = 1.00 \pm 0.03 .
\]

\(^4\) The values for the meson decay constants are from Ref. [17], with the conservative error estimate discussed in [20]. The results for the bag parameters are from [19].
As can be seen, the meson decay constants have errors below the 5% level. For the bag parameters the absolute errors are still at the 10% level, but the error drops to 3% in the ratio, that is sensitive to $SU(3)$ breaking effects only. This is why the ratio $\Delta m_{B_s}/\Delta m_{B_d}$ gives more significant constraint in Fig. 1 with respect to $\Delta m_{B_d}$ only.

The last class of hadronic matrix elements is the one of multi-hadron final states, such as the two-body non-leptonic decays $B \to \pi \pi$ and $B \to K \pi$, as well as many other processes with more than one hadron in the final state. These are the most difficult ones to be estimated from first principles with high accuracy. A lot of progress in the recent pass has been achieved thanks to QCD factorization [21] and the SCET [22] approaches, which provide factorization formuale to relate these hadronic matrix elements to two-body hadronic form factors in the large $m_b$ limit. However, it is fair to say that the errors associated to the $\Lambda_{QCD}/m_b$ corrections are still quite large. This subject is quite interesting by itself, but is beyond the scope of these lectures, where we focus on clean $B$-physics observables for NP studies.

To this purpose, the only interesting non-leptonic channels are those where, with suitable ratios, or using $SU(2)$ relations among hadronic matrix elements, we can eliminate completely all hadronic unknowns. Examples of this type are the $B \to DK$ channels discussed in Section 2.3.

### 2.2 Time evolution of neutral mesons

The non-vanishing amplitude mixing the quasi-stable neutral pseudoscalar mesons ($M^0 \equiv B^0, B^0_d, D^0, \text{or } K^0$) with the corresponding anti mesons induces a time-dependent oscillations between these states. An initially produced $M^0$ or $\bar{M}^0$ evolves in time into a superposition of $M^0$ and $\bar{M}^0$.

For the sake of simplicity, let’s concentrate on the case of $B$ mesons. Denoting by $|B^0(t)\rangle$ (or $|\bar{B}^0(t)\rangle$) the state vector of a $B$ meson which is tagged as a $B^0$ (or $\bar{B}^0$) at time $t=0$, the time evolution of these states is governed by the following equation:

$$
\frac{d}{dt} \begin{pmatrix} |B(t)\rangle \\ \bar{|B(t)\rangle} \end{pmatrix} = \begin{pmatrix} M - i \Gamma/2 \end{pmatrix} \begin{pmatrix} |B(t)\rangle \\ \bar{|B(t)\rangle} \end{pmatrix},
$$

(44)

where the mass-matrix $M$ and the decay-matrix $\Gamma$ are $t$-independent, Hermitian $2 \times 2$ matrices. CPT invariance implies that $M_{11} = M_{22}$ and $\Gamma_{11} = \Gamma_{22}$, while the off-diagonal element $M_{12} = M_{21}$ is the one we can compute using the effective Lagrangian $\mathcal{L}_{\Delta B^0 = 2}$.

The mass eigenstates are the eigenvectors of $M - i \Gamma/2$. We express them in terms of the flavor eigenstates as

$$
|B_L\rangle = p|B^0\rangle + q|\bar{B}^0\rangle, \quad |B_H\rangle = p|B^0\rangle - q|\bar{B}^0\rangle,
$$

(45)

with $|p|^2 + |q|^2 = 1$. Note that, in general, $|B_L\rangle$ and $|B_H\rangle$ are not orthogonal to each other. The time evolution of the mass eigenstates is governed by the two eigenvalues $M_H - i \Gamma_H/2$ and $M_L - i \Gamma_L/2$:

$$
|B_{H,L}(t)\rangle = e^{-(iM_{H,L} + \Gamma_{H,L}/2)t} |B_{H,L}(t=0)\rangle.
$$

(46)

For later convenience it is also useful to define

$$
m = \frac{M_H + M_L}{2}, \quad \Gamma = \frac{\Gamma_L + \Gamma_H}{2}, \quad \Delta m = M_H - M_L, \quad \Delta \Gamma = \Gamma_L - \Gamma_H.
$$

(47)

With these conventions the time evolution of initially tagged $B^0$ or $\bar{B}^0$ states is

$$
|B^0(t)\rangle = e^{-imt} e^{-\Gamma t/2} \begin{bmatrix} f_+(t) |B^0\rangle + \frac{q}{p} f_-(t) |\bar{B}^0\rangle \\ \frac{p}{q} f_-(t) |B^0\rangle + f_+(t) |\bar{B}^0\rangle \end{bmatrix},
$$

$$
|\bar{B}^0(t)\rangle = e^{-imt} e^{-\Gamma t/2} \begin{bmatrix} f_+(t) |\bar{B}^0\rangle + \frac{q}{p} f_-(t) |B^0\rangle \\ \frac{p}{q} f_-(t) |\bar{B}^0\rangle + f_+(t) |B^0\rangle \end{bmatrix},
$$

(48)

where

$$
f_+(t) = \cosh \frac{\Delta m t}{4} \cos \frac{\Delta m t}{2} - i \sinh \frac{\Delta m t}{4} \sin \frac{\Delta m t}{2},
$$

(49)
\[ f_-(t) = -\sinh \frac{\Delta \Gamma t}{4} \cos \frac{\Delta m t}{2} + i \cosh \frac{\Delta \Gamma t}{4} \sin \frac{\Delta m t}{2}, \quad (50) \]

In both \( B_s \) and \( B_d \) systems the following hierarchies holds: \(|\Gamma_{12}| \ll |M_{12}| \) and \( \Delta \Gamma \ll \Delta m \). They are experimentally verified and can be traced back to the fact that \(|\Gamma_{12}| \) is a genuine long-distance \( O(G_F^2) \) effect (it is indeed related to the absorptive part of the box diagrams in Fig. 3) which do not share the large \( m_t \) enhancement of \(|M_{12}| \) (which is a short-distance dominated quantity). Taking into account this hierarchy leads to the following approximate expressions for the quantities appearing in the time-evolution formulae in terms of \( M_{12} \) and \( \Gamma_{12} \):

\[
\Delta m = 2 |M_{12}| \left[ 1 + \mathcal{O} \left( \frac{\Gamma_{12}}{|M_{12}|} \right)^2 \right], \quad (51)
\]

\[
\Delta \Gamma = 2 |\Gamma_{12}| \cos \phi \left[ 1 + \mathcal{O} \left( \frac{\Gamma_{12}}{|M_{12}|} \right)^2 \right], \quad (52)
\]

\[
\frac{q}{p} = -e^{-i\phi_B} \left[ 1 - \frac{1}{2} \frac{\Gamma_{12}}{|M_{12}|} \sin \phi + \mathcal{O} \left( \frac{\Gamma_{12}}{|M_{12}|} \right)^2 \right], \quad (53)
\]

where \( \phi = \text{arg}(-M_{12}/\Gamma_{12}) \) and \( \phi_B \) is the phase of \( M_{12} \). Note that \( \phi_B \) thus defined is not measurable and depends on the phase convention adopted, while \( \phi \) is a phase-convention quantity which can be measured in experiments.

Taking into account the above results, the time-dependent decay rates of an initially tagged \( B^0 \) or \( \bar{B}^0 \) state into some final state \( f \) can be written as

\[
\Gamma[B^0(t = 0) \to f(t)] = N_0 |A_f|^2 e^{-\Gamma t} \left\{ \frac{1 + |\lambda_f|^2}{2} \cosh \frac{\Delta \Gamma t}{2} + \frac{1 - |\lambda_f|^2}{2} \cos(\Delta m t) - \text{Re} \lambda_f \sinh \frac{\Delta \Gamma t}{2} - \text{Im} \lambda_f \sin(\Delta m t) \right\},
\]

\[
\Gamma[\bar{B}^0(t = 0) \to f(t)] = N_0 |A_f|^2 \left( 1 + \left| \frac{\Gamma_{12}}{M_{12}} \right| \sin \phi \right) e^{-\Gamma t} \left\{ \frac{1 + |\lambda_f|^2}{2} \times \cosh \frac{\Delta \Gamma t}{2} - \frac{1 - |\lambda_f|^2}{2} \cos(\Delta m t) - \text{Re} \lambda_f \sinh \frac{\Delta \Gamma t}{2} + \text{Im} \lambda_f \sin(\Delta m t) \right\},
\]

where \( N_0 \) is the flux normalization and, following the standard notation, we have defined

\[
\lambda_f = \frac{q}{p} \frac{\bar{A}_f}{A_f} \approx -e^{-i\phi_B} \frac{\bar{A}_f}{A_f} \left[ 1 - \frac{1}{2} \frac{\Gamma_{12}}{|M_{12}|} \sin \phi \right] \quad (54)
\]

in terms of the decay amplitudes

\[
A_f = \langle f | \mathcal{L}_{\Delta F = 1} | B^0 \rangle, \quad \bar{A}_f = \langle f | \mathcal{L}_{\Delta F = 1} | \bar{B}^0 \rangle. \quad (55)
\]

From the above expressions it is clear that the key quantity we can access experimentally in the time-dependent study of \( B \) decays is the combination \( \lambda_f \). Both real and imaginary parts of \( \lambda_f \) can be measured, and indeed this is a phase-convention independent quantity: the phase convention in \( \phi_B \) is compensated by the phase convention in the decay amplitudes. In other words, what we can measure is the weak-phase difference between \( M_{12} \) and the decay amplitudes.

For generic final states, \( \lambda_f \) is a quantity that is difficult to evaluate. However, it becomes particularly simple in the case where \( f \) is a CP eigenstate, \( \text{CP}|f\rangle = \eta_f|f\rangle \), and the weak phase of the decaying
amplitude is known. In such case $\bar{A}_f/A_f$ is a pure phase factor ($|\bar{A}_f/A_f|=1$), determined by the weak phase of the decaying amplitude:

$$\lambda_f|_{\text{CP-eigen.}} = \eta_f \frac{q}{p} e^{-2i\phi_A}, \quad A_f = |A_f| e^{i\phi_A}, \quad \eta_f = \pm 1.$$  \hspace{1cm} (56)

The most clean example of this type of channels is the $|\psi K_s\rangle$ final state for $B_d$ decays. In this case the final state is a CP eigenstate and the decay amplitude is real (to a very good approximation) in the standard CKM phase convention. Indeed the underlying partonic transition is dominated by the Cabibbo-allowed tree-level process $b \to c\bar{c}s\bar{s}$, which has a vanishing phase in the standard CKM phase convention, and also the leading one-loop corrections (top-quark penguins) have the same vanishing weak phase. Since in the $B_d$ system we can safely neglect $\Gamma_{12}/M_{12}$, this implies

$$\lambda_{\psi K_s}^{B_d} = -e^{-i\phi_{B_d}}, \quad \text{Im} \left( \lambda_{\psi K_s}^{B_d} \right)_{\text{SM}} = \sin(2\beta),$$  \hspace{1cm} (57)

where the SM expression of $\phi_{B_d}$ is nothing but the phase of the CKM combination $(V_{tb}^* V_{td})^2$ appearing in Eq. (37). Given the smallness of $\Delta\Gamma_d$, this quantity is easily extracted by the ratio

$$\frac{\Gamma[\bar{B}_d(t=0) \to \psi K_s(t)] - \Gamma[B^0(t=0) \to f \psi K_s(t)]}{\Gamma[B^0(t=0) \to f \psi K_s(t)]} \approx \frac{\Gamma[\bar{B}_d(t=0) \to f(t)] - \Gamma[B^0(t=0) \to f(t)]}{\Gamma[B^0(t=0) \to f(t)]} = \frac{\Gamma_{12}}{M_{12}} \sin(\Delta m_{B_d} t),$$

which can be considered the golden measurement of $B$ factories.

Another class of interesting final states are CP-conjugate channels $|f\rangle$ and $|\bar{f}\rangle$ which are accessible only to $B^0$ or $\bar{B}^0$ states, such that $|A_f| = |\bar{A}_f|$ and $\bar{A}_f = A_f=0$. Typical examples of this type are the charged semileptonic channels. In this case the asymmetry

$$\frac{\Gamma[\bar{B}^0(t=0) \to f(t)] - \Gamma[B^0(t=0) \to f(t)]}{\Gamma[B^0(t=0) \to f(t)]} = \frac{\Gamma_{12}}{M_{12}} \sin(\Delta m_{B_d} t),$$

turns out to be time-independent and a clean way to determine the indirect CP-violating phase $\phi$.

### 2.2.1 CP violation in $B_s$ mixing

Till very recently the CP violating phase of $B_s$–$B_s$ mixing was the last missing ingredient of down-type $\Delta F = 2$ observables. The golden channel for the measurement of this phase is the time-dependent analysis of the $B_s(B_s) \to \psi\phi$ decay. At the quark level $B_s \to \psi\phi$ share the same virtues of $B_d \to \psi K$ (partonic amplitude of the type $b \to c\bar{c}s\bar{s}$), which is used to extract the phase of $B_d$-$\bar{B}_d$ mixing. However, there a few points which makes this measurement much more challenging:

- The $B_s$ oscillations are much faster ($\Delta m_{B_s}/\Delta m_{B_d} \approx F_{B_s}^2/F_{B_d}^2 |V_{ts}/V_{td}|^2 \approx 30$), making the time-dependent analysis quite difficult (and essentially inaccessible at $B$ factories).

- Contrary to $|\psi K\rangle$, which has a single angular momentum and is a pure CP eigenstate, the vector-vector state $|\psi\phi\rangle$ produced by the $B_s$ decay has different angular momenta, corresponding to different CP eigenstates. These must be disentangled with a proper angular analysis of the final four-body final state $|\ell^+\ell^-\phi(K^+K^-)\rangle$. To avoid contamination from the nearby $|\psi f_0\rangle$ state, the fit should include also a $|\ell^+\ell^-\phi(K^+K^-)_{s-wave}\rangle$ component, for a total of ten independent (and unknown) weak amplitudes.

- Contrary to the $B_d$ system, the width difference cannot be neglected in the $B_s$ case, leading to an additional key parameter to be included in the fit.

Modulo the experimental difficulties listed above, the process is theoretically clean and a complete fit of the decay distributions should allow the extraction of

$$\lambda_{\psi\phi}^{B_s} = -e^{-i\phi_{B_s}},$$  \hspace{1cm} (58)
where the SM prediction is

$$\phi_{B_s}^{SM} = -\arg \left( \frac{(V_{tb}^* V_{ts})^2}{|V_{tb}^* V_{ts}|^2} \right) = -0.04 \pm 0.01 \ . \quad (59)$$

The tiny value of $\phi_{B_s}^{SM}$ implies that, within the SM, no CP asymmetry should be observed in the near future. The present status of the combined fit of $\Delta \Gamma_s$ and $\phi_{B_s}$ as obtained by LHCb and other experiments is shown in Fig. 5. As can be noted, at present these is a good agreement with the SM prediction. However, contrary to all other $\Delta F = 2$ observables, in this case the theory error is still subleading and there is ample room for improving the precision on the experimental side.

As we will see in Sect. 3.1 a clear evidence for $\phi_{B_s} \neq \phi_{B_s}^{SM}$ would not only signal the presence of physics beyond the SM, but would also rule out the whole class of MFV models.

### 2.3 CP violation in charged $B$ decays

Among non-leptonic channels $B^\pm \to DK$ decays are particularly interesting since, via appropriate asymmetries, allows us to extract the CKM angle $\gamma$ in a very clean way. The extraction of $\gamma$ involves only tree-level $B$ decay amplitudes, and is virtually free from hadronic uncertainties (which are eliminated directly by data). It is therefore an essential element for a precise determination of the SM Yukawa couplings also in presence of NP.

The main strategy is based on the following two observations:

- The partonic amplitudes for $B^- \to \bar{D}K^- (b \to c\bar{s}b)$ and $B^- \to \bar{D}K^- (b \to u\bar{c}s)$ are pure tree-level amplitudes (no penguins allowed given the four different quark flavors). As a result, their weak phase difference is completely determined and is $\gamma = \arg \left( -V_{ud} V_{ub}^*/V_{cd} V_{cb}^* \right)$.

- Thanks to $D-\bar{D}$ mixing, there are several final states $f$ accessible to both $D$ and $\bar{D}$, where the two tree-level amplitudes can interfere. By combining the four final states $B^\pm \to fK^\pm$ and $B^\pm \to fK^\mp$, we can extract $\gamma$ and all the relevant hadronic unknowns of the system.$^5$

---

$^5$ The quoted error takes into account possible sub-leading amplitudes contributing to the $B_s \to \psi\phi$ decay with different CKM structure.
The first strategy, proposed by Gronau, London, and Wyler [24] was based on the selection of $D(\bar{D})$ decays to two-body $CP$ eigenstates. But it has later been realized that any final state accessible to both $D$ and $\bar{D}$ (such as the $K^\pm\pi^\mp$ channels [25], or multibody final states [26]) may work as well.

Let’s start from the case of $D(\bar{D})$ decays to CP eigenstates, where the formalism is particularly transparent. The key quantity is the ratio

$$r_B e^{i\delta_B} = \frac{A(B^+ \to D^0 K^+)}{A(B^+ \to D^0 \bar{K}^+)} .$$  \hspace{1cm} (60)

where $\delta$ is a strong phase. Denoting CP-even and CP-odd final states $f_+$ and $f_-$, we then have

\[
\begin{align*}
A(B^- \to f_+ K^-) &= A_0 \times \left[ 1 + r_B e^{i(\delta_B - \gamma)} \right] \\
A(B^- \to f_- K^-) &= A_0 \times \left[ 1 - r_B e^{i(\delta_B - \gamma)} \right] \\
A(B^+ \to f_+ K^+) &= A_0 \times \left[ 1 + r_B e^{i(\delta_B + \gamma)} \right] \\
A(B^+ \to f_- K^+) &= A_0 \times \left[ 1 - r_B e^{i(\delta_B + \gamma)} \right] \hspace{1cm} (61)
\end{align*}
\]

It is clear that combining the four rates we can extract the three hadronic unknowns ($A_0$, $r_B$, and $\gamma$) as well as $\gamma$. It is also clear that the sensitivity to $\gamma$ vanishes in the limit $r_B \to 0$, and indeed the main limitation of this method is that $r_B$ turns out to be very small.

The formalism is essentially unchanged if we consider final states that are not CP eigenstates, such as the $K^\pm\pi^\mp$ states. These have the advantage that the suppression of $r_B$ is partially compensated by the CKM suppression of the corresponding $D(\bar{D}) \to K^\pm\pi^\mp$ decays. Indeed the effective relevant ratio becomes

$$r_{\text{eff}} e^{i\delta_{\text{eff}}} = \frac{A(B^+ \to D^0 K^+)}{A(B^+ \to D^0 \bar{K}^+)} \times \frac{A(D^0 \to K^- \pi^+)}{A(D^0 \to K^- \pi^+)}$$ \hspace{1cm} (62)

which is substantially larger than $r_B$.

Once $r_B$ and $\delta_B$ (or $r_{\text{eff}}$ and $\delta_{\text{eff}}$) are determined from data, the extraction of $\gamma$ has essentially no theoretical uncertainty. In principle a theoretical error could be induced by the neglected CP-violating effects in charm mixing. In practice, the experimental bounds on charm mixing make this effect totally negligible. The key issue is only collecting high statistics on this highly-suppressed decay modes: a clear target for $B$ physics at hadron machines.

### 2.4 Rare FCNC $B$ decays

On general grounds, theoretical predictions for exclusive FCNC decays are not easy: non-perturbative effects are difficult to be kept under good theoretical control. Even if the final state involve only one hadron, in most of the kinematical region re-scattering effects of the type $B \to K^* H \tilde{H} \to K^* \ell^+ \ell^-$ are possible, making difficult to estimate precisely the decay rate.

However, there are a few exceptions. In the $B \to K^* \ell^+ \ell^-$ case the largest source of uncertainty is the normalization of the hadronic form factors. The theoretical error can be substantially reduced in appropriate ratios or differential distributions. A clean example of this type is the normalized forward-backward asymmetry in $B \to K^* \ell^+ \ell^-$. An even cleaner case is the pure leptonic decays, $B\to s, d \to \ell^+ \ell^-$, where re-scattering effects are negligible (due to the peculiar choice of initial and final state) and all relevant non-perturbative effects are encoded into the meson decay constant (that is easily accessible on the Lattice).
2.4.1 The forward-backward asymmetry in $B \rightarrow K^*\ell^+\ell^-$

The observable is defined as

$$A_{FB}(s) = \frac{1}{d\Gamma(B \rightarrow K^*\mu^+\mu^-)/ds} \int_{-1}^{1} d\cos\theta \frac{d^2\Gamma(B \rightarrow K^*\mu^+\mu^-)}{ds d\cos\theta} \text{sgn}(\cos\theta),$$

(63)

where $\theta$ is the angle between the momenta of $\mu^+$ and $\bar{B}$ in the dilepton center-of-mass frame. Assuming that the leptonic current has only a vector ($V$) or axial-vector ($A$) structure (as in the SM), the FB asymmetry provides a direct measure of the $A$–$V$ interference. Indeed, at the lowest-order one can write

$$A_{FB}(q^2) \propto \text{Re} \left\{ C_{10}^A \left[ \frac{q^2}{m_b^2} C_{9V}^{\text{eff}} + r(q^2) \frac{m_b C_{7\gamma}}{m_B} \right] \right\},$$

where $r(q^2)$ is an appropriate ratio of $B \rightarrow K^*$ vector and tensor form factors [27]. There are three main features of this observable that provide a clear and independent short-distance information:

1. The position of the zero ($q_0$) of $A_{FB}(q^2)$ in the low-$q^2$ region (see Fig. 6) [27]: as shown by the detailed analyses in Ref. [28,29], the experimental measurement of $q_0^2$ could allow a determination of $C_7/C_9$ at the 10% level.
2. The sign of $A_{FB}(q^2)$ around the zero. This is fixed unambiguously in terms of the relative sign of $C_{10}$ and $C_9$: within the SM one expects $A_{FB}(q^2 > q_0^2) > 0$ for $|\bar{B}| \equiv |b\bar{d}|$ mesons.
3. The relation $A[\bar{B}]_{FB}(q^2) = -A[B]_{FB}(q^2)$. This follows from the CP-odd structure of $A_{FB}$ and holds at the $10^{-3}$ level within the SM [30], where $C_{10}$ has a negligible CP-violating phase.

The FB asymmetry has been measured recently with high statistics by LHCb [31] (see Fig. 6) and, contrary to previous low-statistics results, the LHCb data are in good agreement with the SM prediction. As can be seen in Fig. 6, the experimental errors are not far from the level of the theoretical uncertainty in the CP-averaged FB asymmetry. However, there is still room for more precise test of the theory in other type of (normalized) differential distributions related to CP asymmetries [32, 33].

2.4.2 $B \rightarrow \ell^+\ell^-$

The purely leptonic decays constitute a special case among exclusive transitions. Within the SM only the axial-current operator, $Q_{10A}$, induces a non-vanishing contribution to these decays. As a result, the short-distance contribution is not diluted by the mixing with four-quark operators. Moreover, the hadronic matrix element involved is the simplest we can consider, namely the $B$-meson decay constant
in Eq. (40). As we have seen, present Lattice errors on \( F_{B_d} \) and \( F_{B_s} \) from lattice QCD are already below 5%, and could further improve in the future.

The price to pay for this theoretically-clean amplitude is a strong helicity suppression for \( \ell = \mu \) (and \( \ell = e \)), or the channels with the best experimental signature. Following the recent theoretical analysis in [20], the theoretical branching ratio of the flavor averaged state (equal mixture of \( B_s \) and \( \bar{B}_s \)) into a muon pair (fully inclusive of soft-photon emission) can be written as

\[
\mathcal{B}(B_s \to \mu^+\mu^-)^{\text{SM}} = 3.235 \times 10^{-9} \times \left( \frac{M_t}{173.2 \text{ GeV}} \right)^{3.07} \left( \frac{F_{B_s}}{227 \text{ MeV}} \right)^2 \left| \frac{V_{tb}^* V_{ts}}{4.05 \times 10^{-2}} \right|^2,
\]

(64)

where in the second line we have explicitly separated the present contribution to the error due to \( F_{B_s} \). As far as the other leptons are concerned, we get

\[
\frac{\mathcal{B}(B_s \to \tau^+\tau^-)^{\text{SM}}}{\mathcal{B}(B_s \to \mu^+\mu^-)^{\text{SM}}} = 215, \quad \frac{\mathcal{B}(B_s \to e^+e^-)^{\text{SM}}}{\mathcal{B}(B_s \to \mu^+\mu^-)^{\text{SM}}} = 2.4 \times 10^{-5}.
\]

(65)

The corresponding \( B_d \) modes are both suppressed by an additional factor \( |V_{td}/V_{ts}|^2 F_{B_d}^2 / F_{B_s}^2 \approx 1/30 \).

As recently pointed out in [34], an important point when comparing the above predictions with experiments is the observation that, at present, experiments extract the \( B_s \) decay rates from a time-integrated distribution. As a result, we cannot access the decay rate of a flavor averaged state (that is what is produced at initial time), but its time-integrated evolution. Due to the non-vanishing width difference \( \Delta \Gamma_s \), this imply a nontrivial correction factor of \( O(10\%) \).

What is presently measured by the LHC experiments is the flavor-averaged time-integrated distribution,

\[
\langle \mathcal{B}(B_s \to f) \rangle_{[t]} = \frac{1}{2} \int_0^t dt' \left[ \Gamma(B_s(t') \to f) + \Gamma(\bar{B}_s(t') \to f) \right],
\]

(66)

where \( \Gamma(B_s(t') \to f) \) denotes the decay distribution, as a function of the proper time \( t' \), of a \( B_s \) flavor eigenstate at initial time (and correspondingly for \( \bar{B}_s \)). Following the discussion in sect. 2.2, let’s define

\[
\Gamma_s = \frac{1}{\tau_{B_s}} = \frac{1}{2} \left( \Gamma_s^H + \Gamma_s^L \right), \quad y_s = \frac{\Gamma_s^L - \Gamma_s^H}{2\Gamma_s} = 0.088 \pm 0.014.
\]

(67)

The time-integrated distribution is related to the flavor-averaged rate at \( t = 0 \) [that is what is predicted in Eq. (64)], by

\[
\langle \mathcal{B}(B_s \to f) \rangle_{[t]} = \kappa^f(t, y_s) \langle \mathcal{B}(B_s \to f) \rangle_{[t=0]} = \kappa^f(t, y_s) \frac{\Gamma(B_s \to f) + \Gamma(\bar{B}_s \to f)}{2\Gamma_s},
\]

(68)

where \( \kappa^f(t, y_s) \) is a model- and channel-dependent correction factor.

For the \( \mu^+\mu^- \) final state (inclusive of bremsstrahlung radiation) the SM expression of the \( \kappa^f(t, y_s) \) factor is [35]

\[
\kappa^{\mu\mu}_{\text{SM}}(t, y_s) = \frac{1}{1 - y_s} \left[ 1 - e^{-t/\tau_{B_s}} \sinh \left( \frac{y_s t}{\tau_{B_s}} \right) - e^{-t/\tau_{B_s}} \cosh \left( \frac{y_s t}{\tau_{B_s}} \right) \right],
\]

where on the second line we have given the SM prediction for the fully integrated branching ratio, that is what we should compare with present experimental data.

The LHCb collaboration has recently reported an evidence of the \( B_s \to \mu^+\mu^- \) [36] decays, reporting the following first measurement of the branching ration:

\[
\langle \mathcal{B}(B_s \to \mu\mu) \rangle_{[t=\infty]}^{\text{exp}} = \left( 3.2^{+1.5}_{-1.2} \right) \times 10^{-9}.
\]

(70)
As can be seen, this result is in good agreement with the SM prediction in Eq. (69). However, the error is still large and correspondingly there is still a sizable region of possible new-physics contributions still to be explored.

The strong helicity suppression and the theoretical cleanness make these modes excellent probes of several new-physics models and, particularly, of scalar FCNC amplitudes. Scalar FCNC operators, such as $\bar{b}_R s_L \mu_L \bar{\mu}_R$, are present within the SM but are negligible because of the smallness of down-type Yukawa couplings. On the other hand, these amplitudes could be non-negligible in models with an extended Higgs sector (see Section 3.1.2). In particular, within the MSSM, where two Higgs doublets are coupled separately to up- and down-type quarks, a sizable enhancement of scalar FCNCs can occur at large $\tan \beta = v_u/v_d$. This effect is very small in non-helicity-suppressed $B$ decays (because of the small Yukawa couplings), but could easily enhance $B \to \ell^+\ell^-$ rates by one order of magnitude. The latter possibility is ruled out by the present bounds on $\mathcal{B}(B_s \to \mu\mu)$, resulting in a significant constraint on such class of models.

An illustration of the possible deviations from the SM predictions in a constrained version of the MSSM (that will be discussed in Section 3.2.2) is shown in Fig. 7. This figure shows that the present search for $B_s \to \mu^+\mu^-$ is complementary to the strong limits already sets on such class of models by the direct searches at ATLAS and CMS. In a long-term perspective, the discovery and the precise measurement of all the accessible $B \to \ell^+\ell^-$ channels is one of the most interesting items of the $B$-physics program at hadron colliders.

2.5 CP violation in the charm system

2.5.1 General considerations

On general grounds, long-distance contributions are usually largely dominant with respect to the short-distance ones in charm mixing and decay amplitudes. This happens is because SM short-distance contributions are not top-mass enhanced as in the $B$ and $K$ systems, and are strongly disfavored by the CKM hierarchy with respect to the dominant transition amplitudes into light quarks. Within the SM the genuine short-distance contributions are suppressed by five powers of the Cabibbo angle. Other contrary, long-distance amplitudes into light quarks can be Cabibbo allowed (i.e. not suppressed by any power of $\lambda$) for partonic transitions of the type $c \to usd$, Cabibbo suppressed [$c \to udd(s\bar{s})$], or at most doubly Cabibbo suppressed [$c \to uds\bar{s}$].

Given this hierarchy of amplitudes, within the SM charm physics does not provide interesting pre-
decisions tests of the CKM mechanism. However, the charm system offers a unique opportunity to explore up-type FCNC amplitudes that maybe significantly enhanced over the SM level in possible extensions of the SM. For instance, as shown in Table 1, very stringent constrains on generic $|\Delta C| = 2$ operators can be derived by the experimental constraints on $D\bar{D}$ mixing. The neutral $D$ system is the latest system of neutral mesons where mixing between the particles and anti-particles has been established. The observation of a non-vanishing amplitude at more than 5$\sigma$ has been reported a few months ago by LHCb collaboration and turns out to be consistent with the (long-distance dominated) SM expectation.

While CP-conserving observables in $D$ decays are largely dominated by long-distance effects, CP-violating observables are typically strongly suppressed within the SM and offer a potentially deeper probe of short-distance dynamics. One of the most interesting recent developments in flavor physics has been the experimental evidence of direct CP violation in two-body Cabibbo-suppressed $D$ decays. An asymmetry close to the 1% level has been announced first by LHCb [39] and soon after confirmed both by CDF [40] and by Belle, although none of the experiments has reached the 5$\sigma$ level. Such a large direct CP asymmetry was not expected within the SM according to pre-LHCb theoretical predictions, and the theoretical interpretation [41] of this result has open an interesting debate that is still in progress.

2.5.2 Standard Model vs. New Physics in $\Delta a_{CP}^{\text{dir}}$

The current experimental world average for the direct CP-violating asymmetry in two-body Cabibbo-suppressed $D$ decays can be summarized as follows

$$\Delta a_{CP}^{\text{dir}} \equiv a_{CP}^{\text{dir}}(D \to K^+K^-) - a_{CP}^{\text{dir}}(D \to \pi^+\pi^-) = (-0.67 \pm 0.16) \%,$$ (71)

where

$$a_{CP}^{\text{dir}}(D \to f) = \frac{\Gamma(D^0 \to f) - \Gamma(\bar{D}^0 \to f)}{\Gamma(D^0 \to f) + \Gamma(\bar{D}^0 \to f)}.$$ (72)

The separate determinations of $a_{CP}^{\text{dir}}(D \to K^+K^-)$ and $a_{CP}^{\text{dir}}(D \to \pi^+\pi^-)$ are affected by larger relative uncertainties and, at present, do not allow to establish a clear evidence of CP-violation in one of the two channels.

In order to be non zero, $\Delta a_{CP}^{\text{dir}}$ requires the interference of two amplitudes with different weak and strong phases. Within the SM, taking into account that one of the two amplitudes is necessarily generated at the one-loop level, this implies the following naive expectation $\Delta a_{CP}^{\text{dir}} = O([V_{cb}V_{ub}/V_{ub}V_{ub}]/\alpha_s/\pi) \sim 10^{-4}$ [41], well below the experimental result in Eq. (71). This has led to extensive speculations in the literature that the measurement of $\Delta a_{CP}^{\text{dir}}$ is a signal of NP. This is a particularly exciting possibility, given that reasonable NP models can be constructed in which all related flavor changing neutral current constraints from $D$ meson mixing are satisfied.

The naive expectation for the SM value of $\Delta a_{CP}^{\text{dir}}$ is based on a perturbative (short-distance) estimate of the loop amplitude with suppressed CKM factors. In fact, there is consensus that a SM explanation for $\Delta a_{CP}^{\text{dir}}$ would have to proceed via a dynamical (long-distance) enhancement of specific hadronic matrix elements, the so-called penguin contractions. The latter are nothing but penguin-type matrix elements that vanish at the tree level, with internal light-quark loops ($s$ and $d$): they cannot be estimated reliably in perturbation theory [42]. The enhancement necessary to explain the observed result is quite large compare to the typical size of non-perturbative effects at the charm scale (the naively suppressed penguin contractions should exceed by a factor 3 to 5 the naively dominant tree-level contractions of the same operators [45]). However, such possibility cannot be excluded from first principles and could even lead to a more coherent picture of available data on two-body Cabibbo-suppressed $D$ decays [43].

On the other hand, a value of $\Delta a_{CP}^{\text{dir}}$ of $O(1\%)$ can naturally be accommodated in well-motivated extensions of the SM. In particular, it fits well in models generating at short distances a sizable CP violating phase for the effective $\Delta C = 1$ chromomagnetic operators (see e.g. [41, 44, 45]). Given this
situation, it is important to identify possible future experimental tests able to distinguish standard vs. non-standard explanations of $\Delta a_{\text{CP}}^{\text{dir}}$.

A general prediction of this class of models, that could be used to test this hypothesis from data, are enhanced direct CP violating (DCPV) asymmetries in radiative decay modes [46] (see also [47, 48]). The first key observation to estimate DCPV asymmetries in radiative decay modes is the strong link between the $\Delta C = 1$ chromomagnetic operator, $Q_{8g}$, and the $\Delta C = 1$ electromagnetic-dipole operator, $Q_{7\gamma}$ [these operators are defined as in (35), with the proper replacement of quark fields: $\{b, s\} \rightarrow \{c, u\}$]. In most explicit NP models, the short-distance Wilson coefficients of these two operators are expected to be similar. Moreover, the two operators undergo a strong model-independent mixing (from QCD) in running down from the electroweak scale to the charm scale. Thus if $\Delta a_{\text{CP}}$ is dominated by NP contributions generated by $Q_{8g}$, we can infer that sizable CP asymmetries should occur also in radiative decays, given the presence of a CP-violating electromagnetic-dipole operator.

The second important ingredient is the observation that in the Cabibbo-suppressed $D \to V\gamma$ decays, where $V$ is a light vector meson ($V = \phi, \rho, \omega$), $Q_{7\gamma}$ has a sizable hadronic matrix element. More explicitly, the short-distance contribution induced by $Q_{7\gamma}$, relative to the total (long-distance) amplitude, is substantially larger with respect to the corresponding relative weight of $Q_{8g}$ in $D \to P^+ P^-$ decays. As a result, DCPV asymmetries in these modes could easily reach the few$\times 3\%$ level in presence of NP. An observation of $|a_{V\gamma}| \gtrsim 3\%$ would be a clear signal of physics beyond the SM, and a clean indication of new CP-violating dynamics associated to dipole operators.

3 Flavor physics beyond the SM: models and predictions

If the physics beyond the SM respects the SM gauge symmetry, as we expect from general arguments, the corrections to low-energy flavor-violating amplitudes can be written in the following general form

$$A(f_i \to f_j + X) = A_0 \left[ \frac{c_{\text{SM}}}{M_W^2} + \frac{c_{\text{NP}}}{\Lambda^2} \right],$$

where $\Lambda$ is the energy scale of the new degrees of freedom. This structure is completely general: the coefficients $c_{\text{SM(NP)}}$ may include appropriate CKM factors and eventually a $\sim 1/(16\pi^2)$ suppression if the amplitude is loop-mediated. Given our ignorance about the $c_{\text{NP}}$, the values of the scale $\Lambda$ probed by present experiments vary over a wide range. However, the general result in Eq. (73) allows us to predict how these bounds will improve with future experiments: the sensitivity on $\Lambda$ scale as $N^{1/4}$, where $N$ is the number of events used to measure the observable. This implies that is not easy to increase substantially the energy reach with indirect NP searches only. Moreover, from Eq. (73) it is also clear that indirect searches can probe NP scales well above the TeV for models where ($c_{\text{SM}} \ll c_{\text{NP}}$), namely models which do not respect the symmetries and the symmetry-breaking pattern of the SM.

The bound on representative $\Delta F = 2$ operators have already been shown in Table 1. As can be seen, for $c_{\text{NP}} = 1$ present data probes very high scales. On the other hand, if we insist with the theoretical prejudice that NP must show up not far from the TeV scale in order to stabilize the Higgs sector, then the new degrees of freedom must have a peculiar flavor structure able to justify the smallness of the effective couplings $c_{\text{NP}}$ for $\Lambda = 1$ TeV.

3.1 The Minimal Flavor Violation hypothesis

The main idea of MFV is that flavor-violating interactions are linked to the known structure of Yukawa couplings also beyond the SM. In a more quantitative way, the MFV construction consists in identifying the flavor symmetry and symmetry-breaking structure of the SM and enforce it also beyond the SM.

The MFV hypothesis consists of two ingredients [49]: (i) a flavor symmetry and (ii) a set of symmetry-breaking terms. The symmetry is noting but the large global symmetry $G_{\text{flavor}}$ of the SM Lagrangian in absence of Yukawa couplings shown in Eq. (4). Since this global symmetry, and particularly
the $SU(3)$ subgroups controlling quark flavor-changing transitions, is already broken within the SM, we cannot promote it to be an exact symmetry of the NP model. Some breaking would appear at the quantum level because of the SM Yukawa interactions. The most restrictive assumption we can make to protect in a consistent way quark-flavor mixing beyond the SM is to assume that $Y_u$ and $Y_d$ are the only sources of flavor symmetry breaking also in the NP model. To implement and interpret this hypothesis in a consistent way, we can assume that $G_q$ is a good symmetry and promote $Y_{u,d}$ to be non-dynamical fields (spurions) with non-trivial transformation properties under $G_q$:

$$Y_u \sim (3, \bar{3}, 1), \quad Y_d \sim (3, 1, \bar{3}). \quad (74)$$

If the breaking of the symmetry occurs at very high energy scales, at low-energies we would only be sensitive to the background values of the $Y$, i.e. to the ordinary SM Yukawa couplings. The role of the Yukawa in breaking the flavor symmetry becomes similar to the role of the Higgs in the the breaking of the gauge symmetry. However, in the case of the Yukawa we don’t know (and we do not attempt to construct) a dynamical model which give rise to this symmetry breaking.

Within the effective-theory approach to physics beyond the SM introduced in Section 1.4, we can say that an effective theory satisfies the criterion of Minimal Flavor Violation in the quark sector if all higher-dimensional operators, constructed from SM and $Y$ fields, are invariant under CP and (formally) under the flavor group $G_q$ [49].

According to this criterion one should in principle consider operators with arbitrary powers of the (dimensionless) Yukawa fields. However, a strong simplification arises by the observation that all the eigenvalues of the Yukawa matrices are small, but for the top one, and that the off-diagonal elements of the CKM matrix are very suppressed. Working in the basis in Eq. (6) we have

$$\left[Y_u(Y_u)^\dagger\right]_i^n \approx y^n_i Y_d^i V_{ij}^*.$$  

As a consequence, in the limit where we neglect light quark masses, the leading $\Delta F = 2$ and $\Delta F = 1$ FCNC amplitudes get exactly the same CKM suppression as in the SM:

$$A(d^i \to d^j)_{\text{MFV}} = (V_{ti}^* V_{tj}) A_{\text{SM}}^{(\Delta F=1)} \left[1 + a_1 \frac{16\pi^2 M_W^2}{\Lambda^2}ight], \quad (76)$$

$$A(M_{ij} - \tilde{M}_{ij})_{\text{MFV}} = (V_{ti}^* V_{tj})^2 A_{\text{SM}}^{(\Delta F=2)} \left[1 + a_2 \frac{16\pi^2 M_W^2}{\Lambda^2}ight]. \quad (77)$$

where the $A_{\text{SM}}^{(i)}$ are the SM loop amplitudes and the $a_i$ are $O(1)$ real parameters. The $a_i$ depend on the specific operator considered but are flavor independent. This implies the same relative correction in $s \to d$, $b \to d$, and $b \to s$ transitions of the same type: a key prediction which can be tested in experiment.

As pointed out in Ref. [50], within the MFV framework several of the constraints used to determine the CKM matrix (and in particular the unitarity triangle) are not affected by NP. In this framework, NP effects are negligible not only in tree-level processes but also in a few clean observables sensitive to loop effects, such as the time-dependent CPV asymmetry in $B_d \to \psi K_{L,S}$. Indeed the structure of the basic flavor-changing coupling in Eq. (77) implies that the weak CPV phase of $B_d \to \psi K_{L,S}$ mixing is arg$[(V_{td}^* V_{tb})^2]$, exactly as in the SM. This construction provides a natural (a posteriori) justification of why no NP effects have been observed in the quark sector: by construction, most of the clean observables measured at $B$ factories are insensitive to NP effects in the MFV framework. A comparison of the CKM fits in the SM and in generic MFV models is shown in Fig. 8. Essentially only $\epsilon_K$ and $\Delta m_{B_d}$ (but not the ratio $\Delta m_{B_s}/\Delta m_{B_d}$) are sensitive to non-standard effects within MFV models.

Given the built-in CKM suppression, the bounds on higher-dimensional operators in the MFV framework turns out to be in the TeV range. This can easily be understood by the discussion in Section 1.4.1: the MFV bounds on operators contributing to $\epsilon_K$ and $\Delta m_{B_d}$ are obtained from Eq. (22)
setting $|c_{ij}| = |y_d^2V_{3i}^*V_{3j}|^2$. In Table 2 we report a few representative examples of the bounds on the higher-dimensional operators in the MFV framework. These bounds are very similar to the bounds on flavor-conserving operators derived by precision electroweak tests. This observation reinforces the conclusion that a deeper study of rare decays is definitely needed in order to clarify the flavor problem: the experimental precision on the clean FCNC observables required to obtain bounds more stringent than those derived from precision electroweak tests (and possibly discover new physics) is typically in the 1% – 10% range.

### 3.1.1 General considerations

The idea that the CKM matrix rules the strength of FCNC transitions also beyond the SM has become a very popular concept in the recent literature and has been implemented and discussed by several authors. It is worth stressing that the CKM matrix represents only one part of the problem: a key role in determining the structure of FCNCs is also played by quark masses, or by the Yukawa eigenvalues. In this respect, the MFV criterion illustrated above provides the maximal protection of FCNCs (or the minimal

---

**Table 2**: Bounds on the scale of new physics (at 95% C.L.) for some representative MFV operators (assuming effective coupling $\pm 1/\Lambda^2$, and considering only one operator at a time), with the corresponding observables used to set the bounds.

<table>
<thead>
<tr>
<th>Operator</th>
<th>Bound on $\Lambda$</th>
<th>Observables</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi^\dagger (D_R Y_u^\dagger Y_u \gamma_{\mu\nu} Q_L) (e F_{\mu\nu})$</td>
<td>6.1 TeV</td>
<td>$B \rightarrow X_s \gamma$, $B \rightarrow X_s \ell^+\ell^-$</td>
</tr>
<tr>
<td>$\frac{1}{2} (Q_L Y_u^\dagger Y_u \gamma_{\mu\nu} Q_L)^2$</td>
<td>5.9 TeV</td>
<td>$\epsilon_K$, $\Delta m_{B_d}$, $\Delta m_{B_s}$</td>
</tr>
<tr>
<td>$\phi^\dagger (D_R Y_u^\dagger Y_u \gamma_{\mu\nu} T^a Q_L) (g_6 G_{\mu\nu})$</td>
<td>3.4 TeV</td>
<td>$B \rightarrow X_s \gamma$, $B \rightarrow X_s \ell^+\ell^-$</td>
</tr>
<tr>
<td>$\left( Q_L Y_u^\dagger Y_u \gamma_{\mu\nu} Q_L \right) (\bar{E}<em>R \gamma</em>{\mu\nu} E_R)$</td>
<td>5.7 TeV</td>
<td>$B_s \rightarrow \mu^+\mu^-$, $B \rightarrow K^* \mu^+\mu^-$</td>
</tr>
<tr>
<td>$i \left( Q_L Y_u^\dagger Y_u \gamma_{\mu\nu} Q_L \right) \phi^\dagger D_{\mu} \phi$</td>
<td>4.1 TeV</td>
<td>$B_s \rightarrow \mu^+\mu^-$, $B \rightarrow K^* \mu^+\mu^-$</td>
</tr>
<tr>
<td>$\left( Q_L Y_u^\dagger Y_u \gamma_{\mu\nu} Q_L \right) \left( \bar{T}<em>L \gamma</em>{\mu\nu} L_L \right)$</td>
<td>5.7 TeV</td>
<td>$B_s \rightarrow \mu^+\mu^-$, $B \rightarrow K^* \mu^+\mu^-$</td>
</tr>
<tr>
<td>$\left( Q_L Y_u^\dagger Y_u \gamma_{\mu\nu} Q_L \right) (e D_{\mu} F_{\mu\nu})$</td>
<td>1.7 TeV</td>
<td>$B \rightarrow K^* \mu^+\mu^-$</td>
</tr>
</tbody>
</table>

---

**Fig. 8**: Fit of the CKM unitarity triangle (in 2008) within the SM (left) and in generic extensions of the SM satisfying the MFV hypothesis (right) [13].
violation of flavor symmetry), since the full structure of Yukawa matrices is preserved. At the same time, this criterion is based on a renormalization-group-invariant symmetry argument. Therefore, it can be implemented independently of any specific hypothesis about the dynamics of the new-physics framework. The only two assumptions are: i) the flavor symmetry and its breaking sources; ii) the number of light degrees of freedom of the theory (identified with the SM fields in the minimal case).

This model-independent structure does not hold in most of the alternative definitions of MFV models that can be found in the literature. For instance, the definition of Ref. [51] (denoted constrained MFV, or CMFV) contains the additional requirement that only the effective FCNC operators which play a significant role within the SM are the only relevant ones also beyond the SM. This condition is realized within weakly coupled theories at the TeV scale with only one light Higgs doublet, such as the MSSM with small $\tan\beta$ and small $\mu$ term. However, it does not hold in other frameworks, such as composite-Higgs models (see e.g. [52–54]) or the MSSM with large $\tan\beta$ and/or large $\mu$ term, whose low-energy phenomenology can still be described using the general MFV criterion discussed above.

Although the MFV seems to be a natural solution to the flavor problem, it should be stressed that we are still very far from having proved the validity of this hypothesis from data. A proof of the MFV hypothesis can be achieved only with a positive evidence of physics beyond the SM exhibiting the flavor-universality pattern (same relative correction in $s \rightarrow d$, $b \rightarrow d$, and $b \rightarrow s$ transitions of the same type) predicted by the MFV assumption. While this goal is quite difficult to be achieved, the MFV framework is quite predictive and thus could easily be falsified. Some of the most interesting predictions which could be tested in the near future are the following:

- No new CPV phases in $B_s$ mixing, hence $|\phi_{B_s}| < 0.05$ from $A_{CP}(B_s \rightarrow \psi\phi)$.
- Ratio of $B_s$ and $B_d$ decays into $\ell^+\ell^-$ pairs determined by the CKM matrix: $B(B_d \rightarrow \ell^+\ell^-)/B(B_s \rightarrow \ell^+\ell^-) \approx |V_{td}/V_{ts}|^2$ (see Fig. 9).
- No new CPV phases in $b \rightarrow s\gamma$, hence vanishingly small CP asymmetries in $B \rightarrow K^*\gamma$ and $B \rightarrow K^*\ell^+\ell^-$. 

Violations of these bounds would not only imply physics beyond the SM, but also a clear signal of new sources of flavor symmetry breaking beyond the Yukawa couplings.
3.1.2 MFV at large \( \tan \beta \)

If the Yukawa Lagrangian contains more than one Higgs field, we can still assume that the Yukawa couplings are the only irreducible breaking sources of \( \mathcal{G}_q \), but we can change their overall normalization. A particularly interesting scenario is the two-Higgs-doublet model where the two Higgses are coupled separately to up- and down-type quarks:

\[
L_{Y}^{2HDM} = Q_L Y_d D_R \bar{\phi} + Q_L Y_u U_R \phi_U + L_L Y_e E_R \phi_D + \text{h.c.} \tag{78}
\]

This Lagrangian is invariant under an extra \( U(1) \) symmetry with respect to the one-Higgs Lagrangian in Eq. (3): a symmetry under which the only charged fields are \( D_R \) and \( E_R \) (charge +1) and \( \phi_D \) (charge −1). This symmetry, denoted \( U_{PQ} \), prevents tree-level FCNCs and implies that \( Y_{u,d} \) are the only sources of \( \mathcal{G}_q \) breaking appearing in the Yukawa interaction (similar to the one-Higgs-doublet scenario). Coherently with the MFV hypothesis, we can then assume that \( Y_{u,d} \) are the only relevant sources of \( \mathcal{G}_q \) breaking appearing in all the low-energy effective operators. This is sufficient to ensure that flavor-mixing is still governed by the CKM matrix, and naturally guarantees a good agreement with present data in the \( \Delta F = 2 \) sector. However, the extra symmetry of the Yukawa interaction allows us to change the overall normalization of \( Y_{u,d} \) with interesting phenomenological consequences in specific rare modes.

The normalization of the Yukawa couplings is controlled by the ratio of the vacuum expectation values of the two Higgs fields, or by the parameter \( \tan \beta = \langle \phi_U \rangle / \langle \phi_D \rangle = v_u / v_d \). Defining the eigenvalues \( \lambda_{a,d} \) as in Eq. (6),

\[
\lambda_u = \frac{1}{v_u} \text{diag}(m_u, m_c, m_t), \\
\lambda_d = \frac{1}{v_d} \text{diag}(m_d, m_s, m_b) = \frac{\tan \beta}{v_u} \text{diag}(m_d, m_s, m_b). \tag{79}
\]

For \( \tan \beta \gg 1 \) the smallness of the \( b \) quark can be attributed to the smallness of \( v_d \) with respect to \( v_u \approx v \), rather than to the smallness of the corresponding Yukawa coupling. As a result, for \( \tan \beta \gg 1 \) we cannot anymore neglect down-type Yukawa couplings. Since the \( b \)-quark Yukawa coupling becomes \( \mathcal{O}(1) \), the large-\( \tan \beta \) regime is particularly interesting for all the helicity-suppressed observables in \( B \) physics (i.e., the observables suppressed within the SM by the smallness of the \( b \)-quark Yukawa coupling).

Another important aspect of this scenario is that the the \( U(1)_{PQ} \) symmetry cannot be exact: it has to be broken at least in the scalar potential in order to avoid the presence of a massless pseudoscalar Higgs boson. Even if the breaking of \( U(1)_{PQ} \) and \( \mathcal{G}_q \) are decoupled, the presence of \( U(1)_{PQ} \) breaking sources can have important implications on the structure of the Yukawa interaction, especially if \( \tan \beta \) is large [57, 58]. We can indeed consider new dimension-four operators such as

\[
\epsilon \tilde{Q}_L \lambda_d D_R \bar{\phi}_U \quad \text{or} \quad \epsilon \tilde{Q}_L \lambda_u \tilde{\lambda}_d D_R \bar{\phi}_U, \tag{80}
\]

where \( \epsilon \) denotes a generic MFV-invariant \( U(1)_{PQ} \)-breaking source. Even if \( \epsilon \ll 1 \), the product \( \epsilon \times \tan \beta \) can be \( \mathcal{O}(1) \), inducing large corrections to the down-type Yukawa sector:

\[
\epsilon \tilde{Q}_L \lambda_d D_R \bar{\phi}_U \xrightarrow{\text{v.e.}} \epsilon \tilde{Q}_L \lambda_d D_R \langle \phi_U \rangle = (\epsilon \times \tan \beta) \tilde{Q}_L \lambda_d D_R \langle \phi_D \rangle. \tag{81}
\]

This is what happens in supersymmetry, where the operators in Eq. (80) are generated at the one-loop level \( [\epsilon \sim 1/(16\pi^2)] \), and the large \( \tan \beta \) solution is particularly welcome in the contest of Grand Unified models [59].

One of the clearest phenomenological consequences is a suppression (typically in the 10 − 50\% range) of the \( B \to \ell\nu \) decay rate with respect to its SM expectation [60]. But the most striking signature could arise from the rare decays \( B_{s,d} \to \ell^+\ell^- \) whose rates could still be significantly different from the corresponding SM expectations. A deviation of both \( B_s \to \ell^+\ell^- \) and \( B_d \to \ell^+\ell^- \) respecting the MFV relation \( \Gamma(B_s \to \ell^+\ell^-) / \Gamma(B_d \to \ell^+\ell^-) \) illustrated in Fig. 9 would be an unambiguous signature of MFV at large \( \tan \beta \) [55, 62].
3.1.3 Beyond the minimal set-up

The breaking of the flavor group $G_q$ and the breaking of the discrete CP symmetry are not necessarily related: generic MFV models can contain flavor-blind (or flavor-universal) phases [53, 63, 64]. Because of the experimental constraints on electric dipole moments (EDMs), which are generally sensitive to such flavor-blind phases [64, 65], in this more general case the bounds on the scale of new physics are substantially higher with respect to the “minimal” case, where the Yukawa couplings are assumed to be the only breaking sources of both symmetries [49].

The correlation of CP-violating effects in the three down-type $\Delta F = 2$ mixing amplitudes ($B_d, s$ and $K$ meson mixing) is a powerful test of possible flavor-blind phases in a MFV framework. At small $\tan \beta$ there is only one relevant $\Delta F = 2$ operator:

$$\left( \bar{Q}_L Y_u Y_u^\dagger Q_L \right)^2.$$  \(82\)

Since this operator is Hermitian, its coupling must be real and no deviations are expected in the $\Delta F = 2$ sector compared to the case without flavor-blind phases. The situation changes if $\tan \beta$ is large. In this case, thanks to the large bottom Yukawa coupling, additional operators with the insertion of $Y_d$ break the universality between $K$ and $B$ systems. In the limit where we can neglect the strange-quark Yukawa coupling, the extra CPV induced by flavor blind phases is equal in $B_s - \bar{B}_s$ and $B_d - \bar{B}_d$ mixing and does not enter $K^0 - \bar{K}^0$ mixing [53].

As stressed above, the MFV expansion relies on the smallness of the off-diagonal elements of the CKM matrix and the hierarchies between the Yukawa eigenvalues. It does not suffer from the fact of $y_t$ (and possibly $y_b$, at large $\tan \beta$) being sizable. As explicitly shown in Eq. (75), the effect of considering high powers in $y_t$ only modify the overall strength of the basic flavor-violating spurion

$$\left( V^\dagger \lambda^2 V \right)_{i \neq j}.$$  \(83\)

An elegant implementation of the MFV hypothesis, taking into account explicitly the special role the diagonal third-generation Yukawa couplings is obtained with a non-linear realization of the flavor symmetry [53,66]. Particularly interesting is the so-called GMFV case, where both $y_t$ and $y_b$ are assumed to be of order one and their effects are re-summed to all orders [53]. As shown in [53], the flavor symmetry group surviving after this resummation and linearly realised (with small breaking terms) is a $U(2)^3 \times U(1)$ group.

Given the smallness of $y_{c,s}/y_t$ and $y_{s,d}/y_d$, as well as the smallness of the off-diagonal elements of the CKM matrix, the phenomenological predictions derived in the GMFV framework are not different from those obtained with the standard MFV expansion in $Y_u$ and $Y_d$, provided the expansion is carried out up to the first non trivial terms. Indeed the difference between the GMFV predictions derived in [53] with respect to those obtained in [49], employing the standard MFV expansion at large $\tan \beta$, can all be attributed to the presence of flavor-blind phases in the GMFV set-up.

3.2 Flavor breaking in the Minimal Supersymmetric extension of the SM

The Minimal Supersymmetric extension of the SM (MSSM) is one of the most well-motivated and definitely the most studied extension of the SM at the TeV scale. For a detailed discussion of this model we refer to the review in Ref. [67] and to the lectures by D. Kazakov at this school. Here we limit our self to analyse some properties of this model relevant to flavor physics.

The particle content of the MSSM consists of the SM gauge and fermion fields plus a scalar partner for each quark and lepton (squarks and sleptons) and a spin-1/2 partner for each gauge field (gauginos). The Higgs sector has two Higgs doublets with the corresponding spin-1/2 partners (higgsinos) and a Yukawa coupling of the type in Eq. (78). While gauge and Yukawa interactions of the model are com-
pletely specified in terms of the corresponding SM couplings, the so-called soft-breaking sector of the theory contains several new free parameters, most of which are related to flavor-violating observables. For instance the $6 \times 6$ mass matrix of the up-type squarks, after the up-type Higgs field gets a vev ($\langle \phi_U \rangle$), has the following structure

$$\hat{M}_U^2 = \begin{pmatrix} \tilde{m}_{Q_L}^2 & A_U \langle \phi_U \rangle \\ A_U^\dagger \langle \phi_U \rangle & \tilde{m}_{U_R}^2 \end{pmatrix} + O(m_Z, m_{\text{top}}) \;,$$

where $\tilde{m}_{Q_L}$, $\tilde{m}_{U_R}$, and $A_U$ are $3 \times 3$ unknown matrices. Indeed the adjective minimal in the MSSM acronyms refers to the particle content of the model but does not specify its flavor structure.

Because of this large number of free parameters, we cannot discuss the implications of the MSSM in flavor physics without specifying in more detail the flavor structure of the model. The versions of the MSSM analysed in the literature range from the so-called Constrained MSSM (CMSSM), where the complete model is specified in terms of only four free parameters (in addition to the SM couplings), to the MSSM without $R$ parity and generic flavor structure, which contains a few hundreds of new free parameters.

Throughout the large amount of work in the past decades it has became clear that the MSSM with generic flavor structure and squarks in the TeV range is not compatible with precision tests in flavor physics. This is true even if we impose $R$ parity, the discrete symmetry which forbids single s-particle production, usually advocated to prevent a too fast proton decay. In this case we have no tree-level FCNC amplitudes, but the loop-induced contributions are still too large compared to the SM ones unless the squarks are highly degenerate or have very small intra-generation mixing angles. This is nothing but a manifestation in the MSSM context of the general flavor problem illustrated in the first lecture.

The flavor problem of the MSSM is an important clue about the underlying mechanism of supersymmetry breaking. On general grounds, mechanisms of SUSY breaking with flavor universality (such as gauge mediation) or with heavy squarks (especially in the case of the first two generations) tends to be favored. However, several options are still open. These range from the very restrictive CMSSM case, which is a special case of MSSM with MFV, to more general scenarios with new small but non-negligible sources of flavor symmetry breaking.

### 3.2.1 Flavor universality, MFV, and RGE in the MSSM

Since the squark fields have well-defined transformation properties under the SM quark-flavor group $G_q$, the MFV hypothesis can easily be implemented in the MSSM framework following the general rules outlined in Sect. 3.1.

We need to consider all possible interactions compatible with i) softly-broken supersymmetry; ii) the breaking of $G_q$ via the spurion fields $Y_{u,d}$. This allows to express the squark mass terms and the trilinear quark-squark-Higgs couplings as follows [49, 68]:

$$\tilde{m}_{Q_L}^2 = \tilde{m}^2 \left( a_1 \mathbb{I} + b_1 Y_u Y_u^\dagger + b_2 Y_d Y_d^\dagger + b_3 Y_d Y_u Y_u^\dagger Y_d^\dagger + \ldots \right) \;,$$

$$\tilde{m}_{U_R}^2 = \tilde{m}^2 \left( a_2 \mathbb{I} + b_4 Y_d Y_d^\dagger Y_u + \ldots \right) \;,$$

$$A_U = A \left( a_3 \mathbb{I} + b_5 Y_d Y_d^\dagger + \ldots \right) Y_d \;,$$

and similarly for the down-type terms. The dimensional parameters $\tilde{m}$ and $A$, expected to be in the range few 100 GeV – 1 TeV, set the overall scale of the soft-breaking terms. In Eq. (85) we have explicitly

\footnote{Supersymmetry must be broken in order to be consistent with observations (we do not observe degenerate spin partners in nature). The soft breaking terms are the most general supersymmetry-breaking terms which preserve the nice ultraviolet properties of the model. They can be divided into two main classes: i) mass terms which break the mass degeneracy of the spin partners (e.g. sfermion or gaugino mass terms); ii) trilinear couplings among the scalar fields of the theory (e.g. sfermion-sfermion-Higgs couplings).}
shown all independent flavor structures which cannot be absorbed into a redefinition of the leading terms (up to tiny contributions quadratic in the Yukawas of the first two families), when \( \tan \beta \) is not too large and the bottom Yukawa coupling is small, the terms quadratic in \( Y_d \) can be dropped.

In a bottom-up approach, the dimensionless coefficients \( a_i \) and \( b_i \) should be considered as free parameters of the model. Note that this structure is renormalization-group invariant: the values of \( a_i \) and \( b_i \) change according to the Renormalization Group (RG) flow, but the general structure of Eq. (85) is unchanged. This is not the case if the \( b_i \) are set to zero, corresponding to the so-called hypothesis of flavor universality. In several explicit mechanism of supersymmetry breaking, the condition of flavor universality holds at some high scale \( M \), such as the scale of Grand Unification in the CMSSM (see below) or the mass-scale of the messenger particles in gauge mediation (see Ref. [69]). In this case non-vanishing \( b_i \sim (1/4\pi)^2 \ln M^2/M_l^2 \) are generated by the RG evolution. As recently pointed out in Ref. [70] the RG flow in the MSSM-MFV framework exhibit quasi infra-red fixed points: even if we start with all the \( b_i = O(1) \) at some high scale, the only non-negligible terms at the TeV scale are those associated to the \( Y_u Y_d \) structures.

If we are interested only in low-energy processes we can integrate out the supersymmetric particles at one loop and project this theory into the general MFV effective theory approach discussed before. In this case the coefficients of the dimension-six effective operators written in terms of SM and Higgs fields are computable in terms of the supersymmetric soft-breaking parameters. The typical effective scale suppressing these operators (assuming an overall coefficient \( 1/\Lambda^2 \)) is \( \Lambda \sim 4\pi \bar{m} \). Since the bounds on \( \Lambda \) within MFV are in the few TeV range, we then conclude that if MFV holds, the present bounds on FCNCs do not exclude squarks in the few hundred GeV mass range, i.e. well within the LHC reach.

### 3.2.2 The CMSSM framework

The CMSSM, also known as mSUGRA, is the supersymmetric extension of the SM with the minimal particle content and the maximal number of universality conditions on the soft-breaking terms. At the scale of Grand Unification (\( M_{\text{GUT}} \sim 10^{16} \) GeV) it is assumed that there are only three independent soft-breaking terms: the universal gaugino mass (\( \tilde{m}_{1/2} \)), the universal trilinear term (\( A \)), and the universal sfermion mass (\( \tilde{m}_0 \)). The model has two additional free parameters in the Higgs sector (the so-called \( \mu \) and \( B \) terms), which control the vacuum expectation values of the two Higgs fields (determined also by the RG running from the unification scale down to the electroweak scale). Imposing the correct \( W \)- and \( Z \)-boson masses allow us to eliminate one of these Higgs-sector parameters, the remaining one is usually chosen to be \( \tan \beta \). As a result, the model is fully specified in terms of the three high-energy parameters \( \{ \tilde{m}_{1/2}, \tilde{m}_0, A \} \), and the low-energy parameter \( \tan \beta \). This constrained version of the MSSM is an example of a SUSY model with MFV. Note, however, that the model is much more constrained than the general MSSM with MFV: in addition to be flavor universal, the soft-breaking terms at the unification scale obey various additional constraints (e.g. in Eq. (85) we have \( a_1 = a_2 = b_1 = 0 \)).

In the MSSM with \( R \) parity we can distinguish five main classes of one-loop diagrams contributing to FCNC and CP violating processes with external down-type quarks. They are distinguished according to the virtual particles running inside the loops: \( W \) and up-quarks (i.e. the leading SM amplitudes), charged-Higgs and up-quarks, charginos and up-squarks, neutralinos and down-squarks, gluinos and down-squarks. Within the CMSSM, the charged-Higgs and chargino exchanges yield the dominant non-standard contributions.

Given the low number of free parameters, the CMSSM is very predictive and phenomenologically constrained by the precision measurements in flavor physics. The most powerful low-energy constraints come from \( B \rightarrow X_s \gamma \), \( B_s \rightarrow \mu^+\mu^- \), and \( B^+ \rightarrow \tau^+\nu \). In particular, as illustrated in Fig. 7 (left), \( B_s \rightarrow \mu^+\mu^- \) does provide a very significant constraint for large values of \( \tan \beta \).

It is worth to stress that as long as we relax the strong universality assumptions of the CMSSM,

---

\(^8\)More precisely, for each choice of \( \{ \tilde{m}_{1/2}, \tilde{m}_0, A, \tan \beta \} \) there is a discrete ambiguity related to the sign of the \( \mu \) term.
the phenomenology of the model can vary substantially. An illustration of this statement is provided by
the the two panels in Fig. 7, where we compare the predictions for $B_s \rightarrow \mu^+\mu^-$ in the CMSSM and in
its minimal variation, the so-called Non-Universal Higgs Mass (NUHM) scenario. In the latter case only
the condition of universality for the soft breaking terms in the Higgs sectors is relaxed, increasing by one
unit the number of free parameters of the model. As can be noted, the difference is substantial (in both
cases all existing constraints are satisfied). This also illustrate how precise data from flavor physics are
essential to discriminate different versions of the MSSM.

### 3.2.3 The mass insertion approximation in the general MSSM

Flavor universality at the GUT scale is not a general property of the MSSM, even if the model is embed-
edded in a Grand Unified Theory. If this assumption is relaxed, new interesting phenomena can occur in
flavor physics. The most general one is the appearance of gluino-mediated one-loop contributions to
FCNC amplitudes [71]

The main problem when going beyond simplifying assumptions, such as flavor universality or
MFV, is the proliferation in the number of free parameters. A useful model-independent parametriza-
tion to describe the new phenomena occurring in the general MSSM with R parity conservation is the so-
called mass insertion (MI) approximation [72]. Selecting a flavor basis for fermion and sfermion states
where all the couplings of these particles to neutral gauginos are flavor diagonal, the new flavor-violating
effects are parametrized in terms of the non-diagonal entries of the sfermion mass matrices. More pre-
cisely, denoting by $\Delta$ the off-diagonal terms in the sfermion mass matrices (i.e. the mass terms relating
sfermions of the same electric charge, but different flavor), the sfermion propagators can be expanded
in terms of $\delta = \Delta/\bar{m}^2$, where $\bar{m}$ is the average sfermion mass. As long as $\Delta$ is significantly smaller
than $\bar{m}^2$ (as suggested by the absence of sizable deviations form the SM), one can truncate the series
to the first term of this expansion and the experimental information concerning FCNC and CP violating
phenomena translates into upper bounds on these $\delta$’s [73].

The major advantage of the MI method is that it is not necessary to perform a full diagonalization
of the sfermion mass matrices, obtaining a substantial simplification in the comparison of flavor-violating
effects in different processes. There exist four type of mass insertions connecting flavors $i$ and $j$ along
a sfermion propagator: $(\Delta_{ij})_{LL}$, $(\Delta_{ij})_{RR}$, $(\Delta_{ij})_{LR}$ and $(\Delta_{ij})_{RL}$. The indexes $L$ and $R$ refer to the
helicity of the fermion partners.

In most cases the leading non-standard amplitude is the gluino-exchange one, which is enhanced
by one or two powers of the ratio $(\alpha_{\text{strong}}/\alpha_{\text{weak}})$ with respect to neutralino- or chargino-mediated am-
plitudes. When analysing the bounds, it is customary to consider one non-vanishing MI at a time, barring
accidental cancellations. This procedure is justified a posteriori by observing that the MI bounds have
typically a strong hierarchy, making the destructive interference among different MIs rather unlikely.
The bound thus obtained from recent measurements in $B$ and $K$ physics are reported in Table 3. The
bounds mainly depend on the gluino and on the average squark mass, scaling as the inverse mass (the
inverse mass square) for bounds derived from $\Delta F = 2$ ($\Delta F = 1$) observables.

The only clear pattern emerging from these bounds is that there is no room for sizable new sources
of flavor-symmetry breaking around the TeV scale. However, it is too early to draw definite conclusions,
especially given we have no positive evidences of supersymmetry so far: the smallness of the bounds
could be due to some approximate symmetry, if the scale of the soft-breaking terms is not far from the
TeV, or it could simply be an indirect indication of a heavy scale for the superpartners.

### 3.3 Flavor protection in models with partial compositeness

So far we have assumed that the suppression of flavor-changing transitions beyond the SM can be at-
tributed to a flavor symmetry, and a specific form of the symmetry-breaking terms. An interesting alter-
native is the possibility of a generic *dynamical suppression* of flavor-changing interactions, related to
Table 3: The phenomenological upper bounds on \((\delta_{ij}^q)_{MM}\) (left) and \((\delta_{ij}^q)_{LR}\) (right), where \(q = u, d\) and \(M = L, R\). The constraints are given for \(m_\tilde{q} = 1\) TeV and \(m_\tilde{g}_{/m_\tilde{q}}^2 = 1\). The bounds are obtained assuming that the phases suppress the imaginary parts by a factor \(\sim 0.3\) (see Ref. [5] for more details).

<table>
<thead>
<tr>
<th>q</th>
<th>ij</th>
<th>((\delta_{ij}^q)_{MM})</th>
<th>q</th>
<th>ij</th>
<th>((\delta_{ij}^q)_{LR})</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>12</td>
<td>0.03</td>
<td>d</td>
<td>13</td>
<td>2 \times 10^{-4}</td>
</tr>
<tr>
<td>d</td>
<td>13</td>
<td>0.2</td>
<td>d</td>
<td>13</td>
<td>0.08</td>
</tr>
<tr>
<td>d</td>
<td>23</td>
<td>0.6</td>
<td>d</td>
<td>23</td>
<td>0.01</td>
</tr>
<tr>
<td>u</td>
<td>12</td>
<td>0.1</td>
<td>u</td>
<td>12</td>
<td>0.02</td>
</tr>
</tbody>
</table>

the weak mixing of the light SM fermions with the new dynamics at the TeV scale. A mechanism of this type is the so-called RS-GIM mechanism occurring in models with a warped extra dimension. In this framework the hierarchy of fermion masses, which is attributed to the different localization of the fermions in the bulk [74], implies that the lightest fermions are those localized far from the infra-red (SM) brane. As a result, the suppression of FCNCs involving light quarks is a consequence of the small overlap of the light fermions with the lightest Kaluza-Klein excitations [75].

As shown in [76] (see also [77]), also the general features of this class of models can be described by means of an effective theory approach. The two main assumptions of this approach are the following:

- There exists a (non-canonical) basis for the SM fermions where their kinetic terms exhibit a rather hierarchical form:
  \[
  \mathcal{L}_{\text{kin}}^{\text{quarks}} = \sum_{\Psi = Q_L, U_R, D_R} \bar{\Psi} Z_{\Psi}^{-2} \Psi \Psi,
  \]
  \[
  Z_{\Psi} = \text{diag}(z_\Psi^{(1)}, z_\Psi^{(2)}, z_\Psi^{(3)}), \quad z_\Psi^{(1)} \ll z_\Psi^{(2)} \ll z_\Psi^{(3)} \lesssim 1.
  \] (86)
- In such basis there is no flavor symmetry and all the flavor-violating interactions, including the Yukawa couplings, are \(O(1)\).

Once the fields are transformed into the canonical basis, the hierarchical kinetic terms act as a distorting lens, through which all interactions are seen as approximately aligned on the magnification axes of the lens. The hierarchical \(z_\Psi^{(i)}\) can be interpreted as the effect of the mixing of an elementary SM-like sector of massless fermions with a corresponding set of heavy composite fermions: the elementary fermions feel the breaking of the electroweak (and flavor) symmetry only via this mixing.

The values of the \(z_\Psi^{(i)}\) can be deduced, up to an overall normalization, from the know structure of the Yukawa couplings, that can be decomposed as follows

\[
Y_{ij}^u \propto z_Q^{(i)} z_U^{(j)}, \quad Y_{ij}^d \propto z_Q^{(i)} z_D^{(j)}.
\] (87)

Inverting such relations we can express the \(z_\Psi^{(i)}\) combinations appearing in the effective couplings of dimension-six operators involving SM fields [e.g. the combination \((z_Q^{(1)} z_Q^{(2)})^2\) for the operator \((\bar{s}_L \gamma_\mu d_L)^2\), etc.] into appropriate powers of quark masses and CKM angles. The resulting suppression of FCNC amplitudes turns out to be quite effective being linked to the hierarchical structure of the SM Yukawa couplings.

As anticipated, this construction provide an effective description of a wide class of models with a warped extra dimension or, equivalently, four-dimensional models with the mixing between a composite and an elementary sector. However, it should be stressed that this mechanism is not a general feature.
of such models: as shown for instance in [54], also in extra-dimensional (partial-composite) models is possible to postulate the existence of additional symmetries and, for instance, recover a MFV structure.

The dynamical mechanism of hierarchical fermion profiles is quite effective in suppressing FCNCs beyond the SM. In particular, it can be shown that all the dimensions-six FCNC left-left operators, such as the $\Delta F = 2$ terms in Eq. (21), have the same parametric suppression as in MFV [76]. However, a residual problem is present in the left-right operators contributing to CP-violating observables in the kaon or charm system. On the one hand, some tuning is need to avoid the bounds from $\epsilon_K$ [78] and $\epsilon'/\epsilon_K$ [79]. On the other hand, in such class of is not difficult to generate a sizable contribution to $\Delta a_{CP}$ able to saturate the present experimental result [77].

Contrary to most of the models discussed before, in this framework no significant NP effects in the $B$ system are expected. Sizable non-standard contributions in the $K$ and $D$ systems could be hidden by the present theoretical uncertainties. As a result, improving our theoretical description of low-energy flavor dynamics could be the tool to reveal the presence of physics beyond the SM.

4 Conclusions
The absence of significant deviations from the SM in quark flavor physics is a key information about any extension of the SM. Only models with a highly non generic flavor structure can both stabilize the electroweak sector and, at the same time, be compatible with flavor observables. In such models we expect new particles within the LHC reach; however, the structure of the new theory cannot be determined using only the high-$p_T$ data from LHC. As illustrated in these lectures, there are still various open questions about the flavor structure of the model that can be addressed only at low energies, and in particular via $B$, $D$ and $K$ decays.

The set of flavor-physics observables to be measured with higher precision, and the rare transitions to be searched for is limited, if we are interested only on physics beyond the SM. But is far from being a small set. As discussed in these lectures, we still have a limited knowledge about CP violation in the $B_s$ and $D$ systems. Despite significant recent progress, new-physics effects could still be hidden in the helicity suppressed $B_{s,d} \rightarrow \ell^+ \ell^-$ decays. Last but not least, a systematic reduction in the determination of the SM Yukawa couplings, such as the determination of $\gamma$ from $B \rightarrow DK$ decays, could possibly reveal non-standard effects in observables that we have already measured well but we are not able yet to predict with corresponding accuracy, such as $\epsilon_K$ or the $B_d$ mixing phase.

Acknowledgments
I wish to thank the organizers of the 2012 European School of High-Energy Physics for the invitation to this interesting school. I’m also grateful to the students and the discussion leaders for stimulating questions and discussions.

References
Abstract
Supersymmetry, a new symmetry that relates bosons and fermions in particle physics, still escapes observation. Search for supersymmetry is one of the main aims of the Large Hadron Collider. The other possible manifestation of supersymmetry is the Dark Matter in the Universe. The present lectures contain a brief introduction to supersymmetry in particle physics. The main notions of supersymmetry are introduced. The supersymmetric extension of the Standard Model – the Minimal Supersymmetric Standard Model – is considered in more detail. Phenomenological features of the Minimal Supersymmetric Standard Model as well as possible experimental signatures of supersymmetry at the Large Hadron Collider are described. The present limits on supersymmetric particles are presented and the allowed region of parameter space of the MSSM is shown.

1 Introduction: what is supersymmetry
Supersymmetry is a boson-fermion symmetry that is aimed to unify all forces in Nature including gravity within a single framework [1–5]. Modern views on supersymmetry in particle physics are based on a string paradigm, though low energy manifestations of supersymmetry (SUSY) can be possibly found at modern colliders and in non-accelerator experiments.

Supersymmetry emerged from attempts to generalize the Poincaré algebra to mix representations with different spin [1]. It happened to be a problematic task due to “no-go” theorems preventing such generalizations [6]. The way out was found by introducing so-called graded Lie algebras, i.e. adding anti-commutators to usual commutators of the Lorentz algebra. Such a generalization, described below, appeared to be the only possible one within the relativistic field theory.

If \( Q \) is a generator of the SUSY algebra, then acting on a boson state it produces a fermion one and vice versa

\[
\bar{Q} |\text{boson}\rangle = |\text{fermion}\rangle, \quad Q |\text{fermion}\rangle = |\text{boson}\rangle.
\]

Since the bosons commute with each other and the fermions anticommute, one immediately finds that the SUSY generators should also anticommute, they must be fermionic, i.e. they must change the spin by a half-odd amount and change the statistics. The key element of the SUSY algebra is

\[
\{Q_\alpha, \bar{Q}_\dot{\alpha}\} = 2 \sigma^\mu_{\alpha\dot{\alpha}} P_\mu
\]  

where \( Q \) and \( \bar{Q} \) are the generators of the supersymmetry transformation and \( P_\mu \) is the generator of translation, the four-momentum.

In what follows we describe the SUSY algebra in more detail and construct its representations which are needed to build the SUSY generalization of the Standard Model (SM) of fundamental interactions. Such a generalization is based on a softly broken SUSY quantum field theory and contains the SM as the low energy theory.

Supersymmetry promises to solve some problems of the Standard Model and of Grand Unified Theories. In what follows we describe supersymmetry as the nearest option for the new physics on the TeV scale.
2 Motivation for SUSY in particle physics

2.1 Unification with gravity

The general idea is a unification of all forces of Nature including quantum gravity. However, the graviton has the spin 2, while other gauge bosons (the photon, gluons, W and Z weak bosons) have the spin 1. Therefore, they correspond to different representations of the Poincaré algebra. To mix them one can use supersymmetry transformations. Starting with the graviton state of the spin 2 and acting by the SUSY generators we get the following chain of states:

\[ \text{spin } 2 \rightarrow \text{spin } \frac{3}{2} \rightarrow \text{spin } 1 \rightarrow \text{spin } \frac{1}{2} \rightarrow \text{spin } 0. \]

Thus, the partial unification of matter (the fermions) with forces (the bosons) naturally arises from an attempt to unify gravity with the other interactions.

Taking infinitesimal transformations \( \delta \epsilon = \epsilon^\alpha Q_\alpha, \quad \bar{\delta} \bar{\epsilon} = \bar{Q}_\alpha \bar{\epsilon}^\alpha \), and using Eq. (1) one gets

\[ \{ \delta \epsilon, \bar{\delta} \bar{\epsilon} \} = 2 (\epsilon \sigma^\mu \bar{\epsilon}) P_\mu, \tag{2} \]

where \( \epsilon, \bar{\epsilon} \) are transformation parameters. Choosing \( \epsilon \) to be local, i.e. the function of the space-time point \( \epsilon = \epsilon(x) \), one finds from Eq. (2) that the anticommutator of two SUSY transformations is a local coordinate translation, and the theory which is invariant under the local coordinate transformation is the General Relativity. Thus, making SUSY local, one naturally obtains the General Relativity, or the theory of gravity, or supergravity [2].

2.2 Unification of gauge couplings

According to the Grand Unification hypothesis, the gauge symmetry increases with the energy [7]. All known interactions are different branches of the unique interaction associated with a simple gauge group. The unification (or splitting) occurs at the high energy. To reach this goal one has to consider how the couplings change with the energy. It is described by renormalization group equations. In the SM the strong and weak couplings associated with the non-Abelian gauge groups decrease with the energy, while the electromagnetic one associated with the Abelian group on the contrary increases. Thus, it is possible that at some energy scale they are equal.

After the precise measurement of the \( SU(3) \times SU(2) \times U(1) \) coupling constants, it has become possible to check the unification numerically. The three coupling constants to be compared are

\[
\begin{align*}
\alpha_1 &= (5/3) g'/2 / (4\pi) = 5\alpha / (3\cos^2 \theta_W), \\
\alpha_2 &= g^2 / (4\pi) = \alpha / \sin^2 \theta_W, \\
\alpha_3 &= g_s^2 / (4\pi)
\end{align*}
\tag{3}
\]

where \( g', g \) and \( g_s \) are the usual \( U(1), SU(2) \) and \( SU(3) \) couplings and \( \alpha \) is the fine structure constant. The factor of \( 5/3 \) in \( \alpha_1 \) has been included for proper normalization of the generators.

In the modified minimal subtraction (\( \overline{\text{MS}} \)) scheme, the world averaged values of the couplings at the \( Z^0 \) energy are obtained from the fit to the LEP and Tevatron data [8]:

\[
\begin{align*}
\alpha^{-1}(M_Z) &= 128.978 \pm 0.027 \\
\sin^2 \theta_{\overline{\text{MS}}} &= 0.23146 \pm 0.00017 \\
\alpha_s &= 0.1184 \pm 0.0031, \tag{4}
\end{align*}
\]

that gives

\[
\begin{align*}
\alpha_1(M_Z) &= 0.017, \\
\alpha_2(M_Z) &= 0.034, \\
\alpha_3(M_Z) &= 0.118 \pm 0.003. \tag{5}
\end{align*}
\]
Unification of the Coupling Constants in the SM and the minimal MSSM

Fig. 1: The evolution of the inverse of the three coupling constants in the Standard Model (left) and in the supersymmetric extension of the SM (MSSM) (right).

Assuming that the SM is valid up to the unification scale, one can then use the known RG equations for the three couplings. In the leading order they are:

\[
\frac{d\tilde{\alpha}_i}{dt} = b_i \tilde{\alpha}_i^2, \quad \tilde{\alpha}_i = \frac{\alpha_i}{4\pi}, \quad t = \log \left( \frac{Q^2}{\mu^2} \right),
\]

where the coefficients for the SM are \(b_i = (41/10, -19/6, -7)\).

The solution to Eqn. (6) is very simple:

\[
\frac{1}{\tilde{\alpha}_i(Q^2)} = \frac{1}{\tilde{\alpha}_i(\mu^2)} - b_i \log \left( \frac{Q^2}{\mu^2} \right).
\]

The result is demonstrated in Fig. 1 showing the evolution of the inverse of the couplings as a function of the logarithm of energy. In this presentation, the evolution becomes a straight line in the first order. The second order corrections are small and do not cause any visible deviation from the straight line. Fig. 1 clearly demonstrates that within the SM the coupling constant unification at a single point is impossible. It is excluded by more than 8 standard deviations. This result means that the unification can only be obtained if the new physics enters between the electroweak and the Planck scales.

In the SUSY case, the slopes of the RG evolution curves are modified. The coefficients \(b_i\) in Eq. (6) now are \(b_i = (33/5, 1, -3)\). The SUSY particles are assumed to contribute effectively to the running of the coupling constants only for the energies above the typical SUSY mass scale. It turns out that within the SUSY model the perfect unification can be obtained as it is shown in Fig. 1. From the fit requiring the unification one finds for the break point \(M_{SUSY}\) and the unification point \(M_{GUT}\) [9]

\[
M_{SUSY} = 10^{3.4\pm0.9\pm0.4} \text{ GeV},
\]

\[
M_{GUT} = 10^{15.8\pm0.3\pm0.1} \text{ GeV},
\]

\[
\alpha_{GUT}^{-1} = 26.3 \pm 1.9 \pm 1.0.
\]

The first error originates from the uncertainty in the coupling constant, while the second one is due to the uncertainty in the mass splitting between the SUSY particles.

This observation was considered as the first “evidence” for supersymmetry, especially since \(M_{SUSY}\) was found in the range preferred by the fine-tuning arguments.
2.3 Solution to the hierarchy problem

The appearance of two different scales $V \gg v$ in the GUT theory, namely, $M_{GUT}$ and $M_W$, leads to a very serious problem which is called the hierarchy problem. There are two aspects of this problem.

The first one is the very existence of the hierarchy. To get the desired spontaneous symmetry breaking pattern, one needs

$$m_H \sim v \sim 10^{2} \text{ GeV} \quad \quad m_\Sigma \sim V \sim 10^{16} \text{ GeV}$$

with

$$m_H - m_\Sigma \sim 10^{-14} \ll 1,$$

where $H$ and $\Sigma$ are the Higgs fields responsible for the spontaneous breaking of the $SU(2)$ and GUT group, respectively. The question arises of how to get so small number in a natural way.

The second aspect of the hierarchy problem is connected with the preservation of the given hierarchy. Even if we choose the hierarchy like in Eq. (9) the radiative corrections will destroy it! To see this, let us consider the radiative correction to the light Higgs mass given by the Feynman diagram shown in Fig. 2.

This correction which is proportional to the mass squared of the heavy particle, obviously, spoils the hierarchy if it is not cancelled. This very accurate cancelation with a precision $\sim 10^{-14}$ needs a fine-tuning of the coupling constants.

The only known way of achieving this kind of cancelation of quadratic terms (also known as the cancelation of the quadratic divergencies) is supersymmetry. Moreover, SUSY automatically cancels the quadratic corrections in all orders of the perturbation theory. This is due to the contributions of superpartners of ordinary particles. The contribution from boson loops cancels those from the fermion ones because of an additional factor $(-1)$ coming from the Fermi statistics, as shown in Fig. 3.

One can see here two types of contribution. The first line is the contribution of the heavy Higgs boson and its superpartner (higgsino). The strength of the interaction is given by the Yukawa coupling constant $\lambda$. The second line represents the gauge interaction proportional to the gauge coupling constant $g$ with the contribution from the heavy gauge boson and its heavy superpartner (gaugino).
In both cases the cancelation of the quadratic terms takes place. This cancelation is true up to the SUSY breaking scale, $M_{\text{SUSY}}$, which should not be very large ($\leq 1 \text{ TeV}$) to make the fine-tuning natural. Indeed, let us take the Higgs boson mass. Requiring for consistency of the perturbation theory that the radiative corrections to the Higgs boson mass do not exceed the mass itself gives

$$\delta M_h^2 \sim g^2 M_{\text{SUSY}}^2 \sim M_h^2.$$  \hspace{1cm} (10)

So, if $M_h \sim 10^2 \text{ GeV}$ and $g \sim 10^{-1}$, one needs $M_{\text{SUSY}} \sim 10^3 \text{ GeV}$ in order that the relation (10) is valid. Thus, we again get the same rough estimate of $M_{\text{SUSY}} \sim 1 \text{ TeV}$ as from the gauge coupling unification above.

That is why it is usually said that supersymmetry solves the hierarchy problem. We show below how SUSY can also explain the origin of the hierarchy.

### 2.4 Astrophysics and cosmology

The shining matter is not the only one in the Universe. Considerable amount of the energy budget consists of the so-called dark matter. The direct evidence for the presence of the dark matter are flat rotation curves of spiral galaxies [10] (see Fig. 4). To explain these curves one has to assume the existence of a galactic halo made of non-shining matter which takes part in the gravitational interaction. The halo has a size more than twice bigger than a visible galaxy. The other manifestation of existence of the dark matter is the so-called gravitational lensing caused by invisible gravitating matter in the sky [11], which leads to the appearance of circular images of distant stars when the light from them passes through the dark matter.

There are two possible types of the dark matter: the hot one, consisting of light relativistic particles and the cold one, consisting of massive weakly interacting particles (WIMPs) [12]. The hot dark matter might consist of neutrinos, however, this has problems with the galaxy formation. As for the cold dark matter, it has no candidates within the SM. At the same time, SUSY provides an excellent candidate for the cold dark matter, namely, the neutralino, the lightest superparticle [13]. It is neutral, heavy, stable and takes part in weak interactions, precisely what is needed for a WIMP.

### 2.5 Integrability and superstrings

Considerable progress in the SUSY field theories in recent years has shown that they possess some remarkable and attractive properties. For instance, the $N = 4$ maximally supersymmetric Yang-Mills
theory has all the features and seems to provide the first integrable model in 4 space-time dimensions. This model, though being unphysical, attracts much attention nowadays. It has no ultraviolet divergences, keeps conformal invariance at the quantum level and seems to provide exact solutions for the amplitudes. Duality of this theory to the string theory in higher dimensions (AdS/CFT correspondence) allows to go beyond the perturbation theory thus revealing the strong coupling regime. This properties distinguish the SYSY theories by their mathematical nature.

Another motivation for supersymmetry follows from even more radical changes of the basic ideas related to the ultimate goal of the construction of the consistent unified theory of everything. At the moment the only viable conception is the superstring theory [14]. In the superstring theory, the strings are considered as the fundamental objects, closed or open, and are nonlocal in their nature. The ordinary particles are considered as string excitation modes. The interactions of the strings, which are local, generate proper interactions of the usual particles, including the gravitational ones.

To be consistent, the string theory should be conformally invariant in a \( D \)-dimensional target space and have a stable vacuum. The first requirement is valid in the classical theory but may be violated by quantum anomalies. The cancelation of the quantum anomalies takes place when the space-time dimension of the target space equals to the critical one which is \( D_c = 26 \) for the bosonic string and \( D_c = 10 \) for the fermionic one.

The second requirement is that the massless string excitations (the particles of the SM) are stable. This assumes the absence of tachyons, the states with the imaginary mass, which can be guaranteed only in the supersymmetric string theories!

The low energy limit of string theories is a kind of supergravity theory which is a local supersymmetric theory. Besides Einstein gravity it contains new interactions and particles, among them the superpartner of a graviton – gravitino, a fermion with spin 3/2. Supergravity itself is not a consistent quantum field theory and is usually treated as an effective theory. It is used in supersymmetric models of particle physics to provide the soft supersymmetry breaking terms.

2.6 Where is SUSY?

After many years of unsuccessful hunt for supersymmetry in particle physics experiments the natural question arises: where is supersymmetry? We try to answer this question describing searches for SUSY at accelerators, in the deep sky with the help of telescopes, and with the help of the underground facilities. It is obvious, that only direct detection of superpartners can convince people in discovery of supersymmetry, however combined information from the sky might give hints to the mass spectra and confirm the SUSY interpretation of the data.

It seems that despite the absence of confirmation supersymmetry stays an unbeatable candidate for physics beyond the Standard Model. The beauty of SUSY lies in the paradigm of unification of all forces of Nature, the ultimate theory of everything. Therefore search for supersymmetry will continue at LHC and perhaps after it.
3 Basics of supersymmetry

3.1 Algebra of SUSY

Combined with the usual Poincaré and internal symmetry algebra the Super-Poincaré Lie algebra contains additional SUSY generators $Q^i_\alpha$ and $\bar{Q}^\dagger_\dot{\alpha}$ [3]

\[
[P_\mu, P_\nu] = 0, \\
[P_\mu, M_{\rho\sigma}] = i (g_{\mu\rho} P_\sigma - g_{\mu\sigma} P_\rho), \\
[M_{\mu\nu}, M_{\rho\sigma}] = i (g_{\mu\rho} M_{\nu\sigma} - g_{\mu\sigma} M_{\nu\rho} - g_{\nu\rho} M_{\mu\sigma} + g_{\nu\sigma} M_{\mu\rho}), \\
[B_\tau, B_\rho] = i C_{\tau\rho} B_t, \\
[B_\tau, P_\mu] = [B_\tau, M_{\rho\sigma}] = 0, \\
[Q^i_\alpha, P_\mu] = [\bar{Q}^{\dagger\beta}_\dot{\alpha}, P_\mu] = 0, \\
[Q^i_\alpha, M_{\rho\sigma}] = \frac{1}{2} (\sigma_{\mu\nu})_i^\rho_\alpha \bar{Q}^{\dagger\beta}_\dot{\alpha}, \\
[Q^{\dagger\beta}_\dot{\alpha}, M_{\mu\nu}] = -\frac{1}{2} \bar{Q}^{\dagger\beta}(\sigma_{\mu\nu})_i^\rho_\alpha, \\
\{Q^i_\alpha, Q^{\dagger\beta}_\dot{\alpha}\} = 2 \delta^{ij}(\sigma^\nu)_{\alpha\dot{\alpha}} P_\mu, \\
\{Q^i_\alpha, Q^{\dagger\beta}_\dot{\alpha}\} = -2 \epsilon_{\alpha\dot{\alpha}} Z^{ij}, \\
\{Q^i_\alpha, Q^{\dagger\beta}_\dot{\alpha}\} = -2 \epsilon_{\dot{\alpha}i} Z^{ij}. \\
\alpha, \dot{\alpha} = 1, 2 \quad i, j = 1, 2, \ldots, N.
\]

Here $P_\mu$ and $M_{\mu\nu}$ are the four-momentum and angular momentum operators, respectively, $B_\tau$ are the internal symmetry generators, $Q^i_\alpha$ and $\bar{Q}^{\dagger\beta}_\dot{\alpha}$ are the spinorial SUSY generators and $Z_{ij}$ are the so-called central charges; $\alpha$, $\dot{\alpha}$ are the spinorial indices. In the simplest case one has one spinor generator $Q_\alpha$ (and the conjugated one $\bar{Q}_\dot{\alpha}$) that corresponds to the ordinary or $N = 1$ supersymmetry. When $N > 1$ one has the extended supersymmetry.

A natural question arises: how many SUSY generators are possible, i.e. what is the value of $N$? To answer this question, consider massless states. Let us start with the ground state labeled by the energy and the helicity, i.e. the projection of the spin on the direction of momenta, and let it be annihilated by $Q_i$

\[
\text{Vacuum} = |E, \lambda\rangle, \quad Q_i |E, \lambda\rangle = 0.
\]

Then one- and many-particle states can be constructed with the help of creation operators as

<table>
<thead>
<tr>
<th>State</th>
<th>Expression</th>
<th># of states</th>
</tr>
</thead>
<tbody>
<tr>
<td>vacuum</td>
<td>$</td>
<td>E, \lambda\rangle$</td>
</tr>
<tr>
<td>1-particle</td>
<td>$Q_i</td>
<td>E, \lambda\rangle =</td>
</tr>
<tr>
<td>2-particle</td>
<td>$Q_i Q_j</td>
<td>E, \lambda\rangle =</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>N-particle</td>
<td>$Q_1 \ldots Q_N</td>
<td>E, \lambda\rangle =</td>
</tr>
</tbody>
</table>

The total # of states is:

\[
\sum_{k=0}^{N} \binom{N}{k} = 2^N = 2^{N-1} \text{ bosons} + 2^{N-1} \text{ fermions}.
\]

The energy $E$ is not changed, since according to (11) the operators $\bar{Q}_{\dot{i}}$ commute with the Hamiltonian.

Thus, one has a sequence of bosonic and fermionic states and the total number of the bosons equals to that of the fermions. This is a generic property of any supersymmetric theory. However, in
CPT invariant theories the number of states is doubled, since CPT transformation changes the sign of the helicity. Hence, in the CPT invariant theories, one has to add the states with the opposite helicity to the above mentioned ones.

Let us consider some examples. We take $N = 1$ and $\lambda = 0$. Then one has the following set of states:

<table>
<thead>
<tr>
<th>$N = 1$</th>
<th>$\lambda = 0$</th>
<th>helicity</th>
<th>$0 - \frac{1}{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td># of states</td>
<td>1 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hence, the complete $N = 1$ multiplet is

<table>
<thead>
<tr>
<th>$N = 1$</th>
<th>helicity</th>
<th>$-\frac{1}{2}$ 0 $\frac{1}{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td># of states</td>
<td>1 2 1</td>
<td></td>
</tr>
</tbody>
</table>

which contains one complex scalar and one spinor with two helicity states.

This is an example of the so-called self-conjugated multiplet. There are also the self-conjugated multiplets with $N > 1$ corresponding to the extended supersymmetry. Two particular examples are the $N = 4$ super Yang-Mills multiplet and the $N = 8$ supergravity multiplet

<table>
<thead>
<tr>
<th>$N = 4$</th>
<th>SUSY YM</th>
<th>$\lambda = -1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>helicity</td>
<td>$-1 -\frac{1}{2}$ 0 $\frac{1}{2}$ 1</td>
<td></td>
</tr>
<tr>
<td># of states</td>
<td>1 4 6 4 1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N = 8$</th>
<th>SUGRA</th>
<th>$\lambda = -2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-2 -\frac{3}{2}$ $-1 -\frac{1}{2}$ 0 $\frac{1}{2}$ 1 $\frac{3}{2}$ 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 8 28 56 70 56 28 8 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

One can see that the multiplets of extended supersymmetry are very rich and contain a vast number of particles.

The constraint on the number of the SUSY generators comes from the requirement of consistency of the corresponding QFT. The number of supersymmetries and the maximal spin of the particle in the multiplet are related by

$$N \leq 4S,$$

where $S$ is the maximal spin. Since the theories with the spin greater than 1 are non-renormalizable and the theories with the spin greater than 5/2 have no consistent coupling to gravity, this imposes a constraint on the number of the SUSY generators

$$N \leq 4 \quad \text{for renormalizable theories (YM)},$$
$$N \leq 8 \quad \text{for (super)gravity}.$$

In what follows, we shall consider the simple supersymmetry, or the $N = 1$ supersymmetry, contrary to the extended supersymmetries with $N > 1$. In this case, one has the following types of the supermultiplets which are used for the construction of the SUSY generalization of the SM

$$\begin{align*}
(\phi, \psi) & \quad Spin = 0, \quad Spin = 1/2, \\
(\lambda, A_\mu) & \quad Spin = 1/2, \quad Spin = 1
\end{align*}$$

scalar chiral Majorana vector

fermion fermion

each of them contains two physical states, one boson and one fermion. They are called chiral and vector multiplets, respectively. To construct the generalization of the SM one has to put all the particles into these multiplets. For instance, the quarks should go into the chiral multiplet and the photon into the vector multiplet.
3.2 Superspace and supermultiplets

An elegant formulation of the supersymmetry transformations and invariants can be achieved in the framework of the superspace formalism \[4\]. The superspace differs from the ordinary Euclidean (Minkowski) space by adding two new coordinates, \( \theta_\alpha \) and \( \bar{\theta}^\dot{\alpha} \), which are Grassmannian, i.e. anti-commuting, variables

\[
\{ \theta_\alpha, \theta_\beta \} = 0, \quad \{ \bar{\theta}^\dot{\alpha}, \bar{\theta}^\dot{\beta} \} = 0, \quad \theta_\alpha^2 = 0, \quad \bar{\theta}^\dot{\alpha}^2 = 0, \quad \alpha, \beta, \dot{\alpha}, \dot{\beta} = 1, 2.
\]

Thus, we go from the space to the superspace

\[
\text{Space} \quad \Rightarrow \quad \text{Superspace}
\]

\[
x_\mu \quad \Rightarrow \quad x_\mu, \theta_\alpha, \bar{\theta}^\dot{\alpha}
\]

A SUSY group element can be constructed in the superspace in the same way as the ordinary translation in the usual space

\[
G(x, \theta, \bar{\theta}) = e^{i(-x^\mu P_\mu + \theta Q + \bar{\theta} \bar{Q})}.
\]

It leads to a supertranslation in the superspace

\[
x_\mu \rightarrow x_\mu + i \theta \sigma_\mu \bar{\varepsilon} - i \varepsilon \sigma_\mu \bar{\theta},
\theta \rightarrow \theta + \varepsilon,
\bar{\theta} \rightarrow \bar{\theta} + \bar{\varepsilon},
\]

where \( \varepsilon \) and \( \bar{\varepsilon} \) are the Grassmannian transformation parameters. From Eqn. (13) one can easily obtain the representation for the supercharges (11) acting on the superspace

\[
Q_\alpha = \frac{\partial}{\partial \theta_\alpha} - i \sigma^\mu_{\alpha \dot{\alpha}} \bar{\theta}^\dot{\alpha} \partial_\mu, \quad \bar{Q}_{\dot{\alpha}} = \frac{\partial}{\partial \bar{\theta}^{\dot{\alpha}}} + i \theta_\alpha \sigma^\mu_{\alpha \dot{\alpha}} \partial_\mu.
\]

To define the fields on the superspace, consider the representations of the Super-Poincaré group (11) \[3\]. The simplest \( N = 1 \) SUSY multiplets that we discussed earlier are: the chiral one \( \Phi(y, \theta) \) (\( y = x + i \theta \sigma \bar{\theta} \)) and the vector one \( V(x, \theta, \bar{\theta}) \). Being expanded in the Taylor series over the Grassmannian variables \( \theta \) and \( \bar{\theta} \) they give:

\[
\Phi(y, \theta) = A(y) + \sqrt{2} \theta \psi(y) + \theta \theta F(y) =
\]

\[
= A(x) + i \theta \sigma^\mu \partial_\mu A(x) + \frac{1}{4} \theta \theta \bar{\theta} \bar{\theta} \Box A(x)
\]

\[
+ \sqrt{2} \theta \psi(x) - \frac{i}{\sqrt{2}} \theta \theta \partial_\mu \psi(x) \sigma^\mu \bar{\theta} + \theta \theta F(x).
\]

The coefficients are the ordinary functions of \( x \) being the usual fields. They are called the components of the superfield. In Eq. (15) one has 2 bosonic (the complex scalar field \( A \)) and 2 fermionic (the Weyl spinor field \( \psi \)) degrees of freedom. The component fields \( A \) and \( \psi \) are called the superpartners. The field \( F \) is an auxiliary field, it has the "wrong" dimension and has no physical meaning. It is needed to close the algebra (11). One can get rid of the auxiliary fields with the help of equations of motion.

Thus, the superfield contains an equal number of the bosonic and fermionic degrees of freedom. Under the SUSY transformation they convert one into another

\[
\delta_\varepsilon A = \sqrt{2} \varepsilon \psi,
\delta_\varepsilon \psi = i \sqrt{2} \sigma^\mu \bar{\varepsilon} \partial_\mu A + \sqrt{2} \varepsilon F,
\delta_\varepsilon F = i \sqrt{2} \varepsilon \sigma^\mu \partial_\mu \psi.
\]
Notice that the variation of the $F$-component is a total derivative, i.e. it vanishes when integrated over the space-time.

The vector superfield is real $V = V^\dagger$. It has the following Grassmannian expansion:

$$
V(x, \theta, \bar{\theta}) = C(x) + i \theta \chi(x) - i \bar{\theta} \bar{\chi}(x) + \frac{i}{2} \theta \theta [M(x) + iN(x)]
$$

$$
- \frac{i}{2} \bar{\theta} \bar{\theta} [\lambda(x) + \frac{i}{2} \sigma^\mu \partial_\mu \chi(x)] + \frac{1}{2} \theta \theta \bar{\theta} \bar{\theta} [D(x) + \frac{1}{2} \Box C(x)].
$$

The physical degrees of freedom corresponding to the real vector superfield $V$ are the vector gauge field $v_\mu$ and the Majorana spinor field $\lambda$. All other components are unphysical and can be eliminated. Indeed, one can choose a gauge (the Wess-Zumino gauge) where $C = \chi = M = N = 0$, leaving one with only physical degrees of freedom except for the auxiliary field $D$. In this gauge

$$
V = -\theta \sigma^\mu \bar{v}_\mu(x) + i \theta \theta \bar{\theta} \lambda(x) - i \bar{\theta} \bar{\theta} \lambda(x) + \frac{1}{2} \theta \theta \bar{\theta} \bar{\theta} D(x),
$$

$$
V^2 = -\frac{1}{2} \theta \theta \bar{\theta} \bar{\theta} v_\mu(x) v^\mu(x),
$$

$$
V^3 = 0, \quad \text{etc.}
$$

One can define also a field strength tensor (as the analog of $F_{\mu\nu}$ in the gauge theories)

$$
W_\alpha = -\frac{1}{4} \bar{D}^2 V D_\alpha e^{-V}, \quad \bar{W}_\alpha = -\frac{1}{4} D^2 e^V \bar{D}_\alpha e^{-V},
$$

Here $D$ and $\bar{D}$ are the supercovariant derivatives. The field strength tensor in the chosen Wess-Zumino gauge is a polynomial over the component fields:

$$
W_\alpha = T^a \left(-i \lambda_\alpha^a + \theta_\alpha D^a - \frac{i}{2} (\sigma^\mu \bar{\sigma}^\nu) \theta_\alpha F^a_{\mu\nu} + \theta^2 (\sigma^\mu D^a \bar{\lambda}_\alpha^a)\right),
$$

where

$$
F^a_{\mu\nu} = \partial_\mu v_\nu^a - \partial_\nu v_\mu^a + f^{abc} v_\mu^b v_\nu^c, \quad D_{\mu} \bar{\lambda}_\alpha^a = \partial \bar{\lambda}_\alpha^a + f^{abc} v_\mu^b \bar{\lambda}_\alpha^c.
$$

In the Abelian case Eqs. (19) are simplified and take the form

$$
W_\alpha = -\frac{1}{4} \bar{D}^2 D_\alpha V, \quad \bar{W}_\alpha = -\frac{1}{4} D^2 \bar{D}_\alpha V.
$$

### 3.2.1 Construction of SUSY Lagrangians

Let us start with the Lagrangian which has no local gauge invariance. In the superfield notation the SUSY invariant Lagrangians are the polynomials of the superfields. In the same way, as the ordinary action is the integral over the space-time of the Lagrangian density, in the supersymmetric case the action is the integral over the superspace. The space-time Lagrangian density is [3, 4]

$$
\mathcal{L} = \int d^2 \theta \ d^2 \bar{\theta} \Phi_i^+ \Phi_i + \int d^2 \theta \left[ \lambda_i \Phi_i + \frac{1}{2} m_{ij} \Phi_i \Phi_j + \frac{1}{3} y_{ijk} \Phi_i \Phi_j \Phi_k \right] + h.c.
$$

where the first part is the kinetic term and the second one is the superpotential $W$. We use here the integration over the superspace according to the rules of the Grassmannian integration [15]

$$
\int d\theta_\alpha = 0, \quad \int \theta_\alpha d\theta_\beta = \delta_{\alpha\beta}.
$$
Performing the explicit integration over the Grassmannian parameters, we get from Eq. (21)
\[ \mathcal{L} = i \partial_\mu \bar{\psi}_i \bar{\sigma}_\mu \psi_i + A^*_i \Box A_i + F^\dagger_i F_i \]
\[ + \left[ \lambda_i F_i + m_{ij} \left( A_i F_j - \frac{1}{2} \psi_i \psi_j \right) + y_{ijk} \left( A_i A_j F_k - \psi_i \psi_j A_k \right) + h.c. \right]. \]  
(22)

The last two terms are the interaction ones. To obtain the familiar form of the Lagrangian, we have to solve the constraints
\[ \frac{\partial \mathcal{L}}{\partial F^*_k} = F_k + \lambda_k^* + m^*_{ik} A_i^* + y_{ijk}^* A_j^* A_k = 0, \]
\[ \frac{\partial \mathcal{L}}{\partial F_k} = F_k^* + \lambda_k + m_{ik} A_i + y_{ijk} A_i A_j = 0. \]  
(23)

Expressing the auxiliary fields \( F \) and \( F^* \) from these equations, one finally gets
\[ \mathcal{L} = i \partial_\mu \bar{\psi}_i \bar{\sigma}_\mu \psi_i + A^*_i \Box A_i + \frac{i}{2} m_{ij} \psi_i \psi_j - \frac{1}{2} m^*_{ij} \bar{\psi}_i \bar{\psi}_j \]
\[ - y_{ijk} \psi_i \psi_j A_k - y_{ijk}^* \bar{\psi}_i \bar{\psi}_j A_k^* - V(A_i, A_j), \]  
where the scalar potential \( V = F^*_k F_k \). We will return to the discussion of the form of the scalar potential in the SUSY theories later.

Consider now the gauge invariant SUSY Lagrangians. They should contain the gauge invariant interaction of the matter fields with the gauge ones and the kinetic term and the self-interaction of the gauge fields.

Let us start with the gauge field kinetic terms. In the Wess-Zumino gauge one has
\[ W^\alpha W_\alpha |_{\theta 0} = -2i \lambda \sigma^\mu D_\mu \bar{\lambda} - \frac{1}{2} F_{\mu \nu} F^{\mu \nu} + \frac{1}{2} D^2 + \frac{i}{4} F_{\mu \nu} F_{\rho \sigma} \epsilon_{\mu \nu \rho \sigma}, \]  
(25)
where \( D_\mu \bar{\lambda} = \partial_\mu + ig [\gamma_\mu, \bar{\lambda}] \) is the usual covariant derivative and the last, the so-called topological \( \theta \)-term, is the total derivative. The gauge invariant Lagrangian now has the familiar form
\[ \mathcal{L} = \frac{1}{4} \int d^2 \theta W^\alpha W_\alpha + \frac{1}{4} \int d^2 \bar{\theta} \bar{W}^\alpha \bar{W}_\alpha \]
\[ = \frac{1}{2} D^2 - \frac{1}{4} F_{\mu \nu} F^{\mu \nu} - i \lambda \sigma^\mu D_\mu \bar{\lambda}. \]  
(26)

To obtain the gauge-invariant interaction with the matter chiral superfields, one has to modify the kinetic term by inserting the bridge operator
\[ \Phi_i^+ \Phi_i \implies \Phi_i^+ e^{\theta V} \Phi_i. \]  
(27)

The complete SUSY and gauge invariant Lagrangian then looks like
\[ \mathcal{L}_{SUSYYM} = \frac{1}{4} \int d^2 \theta \text{Tr}(W^\alpha W_\alpha) + \frac{1}{4} \int d^2 \bar{\theta} \text{Tr}(\bar{W}^\alpha \bar{W}_\alpha) \]
\[ + \int d^2 \theta d^2 \bar{\theta} \Phi_i \bar{\Phi}_i (e^{\theta V})^\dagger_\beta \Phi^\dagger_\beta \Phi_i + \int d^2 \theta W(\Phi_i) + \int d^2 \bar{\theta} \bar{W}(\bar{\Phi}_i), \]  
(28)
where \( W \) is the superpotential, which should be invariant under the group of symmetry of the particular model. In terms of the component fields the above Lagrangian takes the form
\[ \mathcal{L}_{SUSYYM} = -\frac{1}{4} F_{\mu \nu}^a F^{a \mu \nu} - i \lambda^a \sigma^\mu D_\mu \bar{\lambda}^a + \frac{1}{2} D^a D^\dagger_a \]
\[ + (\partial_\mu A_i - igv_\mu^a T^a A_i) \frac{1}{2} (\partial_\mu A_i - igv_\mu^a T^a A_i) - i \bar{\psi}_i \bar{\sigma}^\mu (\partial_\mu \psi_i - igv_\mu^a T^a \psi_i) \]
\[ - D^a A^\dagger_i T^a A_i - i \sqrt{2} A^\dagger_i T^a \lambda^a \psi_i + i \sqrt{2} \bar{\psi}_i T^a A_i \bar{\lambda}^a + F^\dagger_i F_i \]
\[ + \frac{\partial W}{\partial A_i} F_i + \frac{\partial \bar{W}}{\partial A^\dagger_i} F^\dagger_i \frac{1}{2} \left( \frac{\partial^2 W}{\partial A_i \partial A_j} \psi_i \psi_j - \frac{1}{2} \frac{\partial^2 \bar{W}}{\partial A^\dagger_i \partial A^\dagger_j} \bar{\psi}_i \bar{\psi}_j \right). \]  
(29)
Integrating out the auxiliary fields $D^a$ and $F_i$, one reproduces the usual Lagrangian.

### 3.2.2 The scalar potential

Contrary to the SM, where the scalar potential is arbitrary and is defined only by the requirement of the gauge invariance, in the supersymmetric theories it is completely defined by the superpotential. It consists of the contributions from the $D$-terms and $F$-terms. The kinetic energy of the gauge fields (recall Eq. (26) yields the $\frac{1}{2} D^a D^a$ term, and the matter-gauge interaction (recall Eq. (29) yields the $g D^a T_{ij}^a A_i^* A_j$ one. Together they give

$$L_D = \frac{1}{2} D^a D^a + g D^a T_{ij}^a A_i^* A_j. \quad (30)$$

The equation of motion reads

$$D^a = -g T_{ij}^a A_i^* A_j. \quad (31)$$

Substituting it back into Eq. (30) yields the $D$-term part of the potential

$$L_D = -\frac{1}{2} D^a D^a \implies V_D = \frac{1}{2} D^a D^a, \quad (32)$$

where $D$ is given by Eqn. (31).

The $F$-term contribution can be derived from the matter field self-interaction (22). For a general type superpotential $W$ one has

$$L_F = F_i^* F_i + \left( \frac{\partial W}{\partial A_i} F_i + h.c. \right). \quad (33)$$

Using the equations of motion for the auxiliary field $F_i$

$$F_i^* = -\frac{\partial W}{\partial A_i} \quad (34)$$

yields

$$L_F = -F_i^* F_i \implies V_F = F_i^* F_i, \quad (35)$$

where $F$ is given by Eq. (34). The full scalar potential is the sum of the two contributions

$$V = V_D + V_F. \quad (36)$$

Thus, the form of the Lagrangian is practically fixed by the symmetry requirements. The only freedom is the field content, the value of the gauge coupling $g$, Yukawa couplings $y_{ijk}$ and the masses. Because of the renormalizability constraint $V \leq A^4$ the superpotential should be limited by $W \leq \Phi^3$ as in Eq. (21). All members of the supermultiplet have the same masses, i.e. the bosons and the fermions are degenerate in masses. This property of the SUSY theories contradicts to the phenomenology and requires supersymmetry breaking.

### 4 SUSY generalization of the Standard Model. The MSSM

As has been already mentioned, in the SUSY theories the number of the bosonic degrees of freedom equals that of fermionic. At the same time, in the SM one has 28 bosonic and 90 fermionic degrees of freedom (with the massless neutrino, otherwise 96). So the SM is to a great extent non-supersymmetric. Trying to add some new particles to supersymmetrize the SM, one should take into account the following observations:

- There are no fermions with quantum numbers of the gauge bosons;
– Higgs fields have nonzero vacuum expectation values; hence, they cannot be the superpartners of the quarks and leptons, since this would induce a spontaneous violation of the baryon and lepton numbers;
– One needs at least two complex chiral Higgs multiplets in order to give masses to the up and down quarks.

The latter is due to the form of the superpotential and the chirality of the matter superfields. Indeed, the superpotential should be invariant under the $SU(3) \times SU(2) \times U(1)$ gauge group. If one looks at the Yukawa interaction in the Standard Model, one finds that it is indeed $U(1)$ invariant since the sum of hypercharges in each vertex equals zero. For the up quarks this is achieved by taking the conjugated Higgs doublet $\tilde{H} = i\tau_2 H^\dagger$ instead of $H$. However, in SUSY $H$ is the chiral superfield and hence the superpotential which is constructed out of the chiral fields, may contain only $H$ but not $\tilde{H}$ which is the antichiral superfield.

Another reason for the second Higgs doublet is related to chiral anomalies. It is known that the chiral anomalies spoil the gauge invariance and, hence, the renormalizability of the theory. They are canceled in the SM between the quarks and leptons in each generation [16]

\[ \text{Tr} Y^3 = 3 \times \left( \frac{1}{27} + \frac{1}{27} - \frac{64}{27} \right) - 1 - 1 + 8 = 0 \]

However, if one introduces the chiral Higgs superfield, it contains higgsinos, which are the chiral fermions, and contain the anomalies. To cancel them one has to add the second Higgs doublet with the opposite hypercharge. Therefore, the Higgs sector in the SUSY models is inevitably enlarged, it contains an even number of the Higgs doublets.

**Conclusion:** In the SUSY models the supersymmetry associates the known bosons with the new fermions and the known fermions with the new bosons.

### 4.1 The field content

Consider the particle content of the Minimal Supersymmetric Standard Model [17–19]. According to the previous discussion, in the minimal version we double the number of particles (introducing the superpartner to each particle) and add another Higgs doublet (with its superpartner). Thus, the characteristic feature of any supersymmetric generalization of the SM is the presence of the superpartners (see Fig. 5) [20]. If the supersymmetry is exact, the superpartners of the ordinary particles should have the same masses and have to be observed. The absence of them at modern energies is believed to be explained by the fact that they are very heavy, that means that the supersymmetry should be broken. Hence, if the energy of accelerators is high enough, the superpartners will be created. The particle content of the MSSM then appears as shown in Table 1. Hereafter, a tilde denotes the superpartner of the ordinary particle.

The presence of the extra Higgs doublet in the SUSY model is a novel feature of the theory. In the MSSM one has two doublets with the quantum numbers $(1,2,-1)$ and $(1,2,1)$, respectively:

\[
H_1 = \begin{pmatrix} H_1^0 \\ H_1^- \end{pmatrix} = \begin{pmatrix} v_1 + \frac{S_1 + iP_1}{\sqrt{2}} \\ H_1^- \end{pmatrix},
\]

\[
H_2 = \begin{pmatrix} H_2^+ \\ H_2^0 \end{pmatrix} = \begin{pmatrix} H_2^+ \\ v_2 + \frac{S_2 + iP_2}{\sqrt{2}} \end{pmatrix},
\]

where \(v_i\) are the vacuum expectation values of the neutral components of the Higgs doublets.

Hence, one has \(8 = 4 + 4 = 5 + 3\) degrees of freedom. As in the case of the SM, 3 degrees of freedom can be gauged away, and one is left with 5 physical states compared to 1 in the SM. Thus, in
the MSSM, as actually in any two Higgs doublet model, one has five physical Higgs bosons: two \(CP\)-even neutral Higgs, one \(CP\)-odd neutral Higgs and two charged ones. We consider the mass eigenstates below.

4.2 Lagrangian of the MSSM

Now we can construct the Lagrangian of the MSSM. It consists of two parts; the first part is the SUSY generalization of the Standard Model, while the second one represents the SUSY breaking as mentioned above.

\[ \mathcal{L}_{MSSM} = \mathcal{L}_{SUSY} + \mathcal{L}_{Breaking}, \]  

(37)

where

\[ \mathcal{L}_{SUSY} = \mathcal{L}_{Gauge} + \mathcal{L}_{Yukawa}. \]  

(38)

We will not describe the gauge part here, since it is essentially the gauge invariant kinetic terms, but rather concentrate on Yukawa terms. They are given by the superpotential which is nothing else but the usual Yukawa terms of the SM with the fields replaced by the superfields as explained above.

\[ \mathcal{L}_{Yukawa} = \epsilon_{ij} \left( y_{\text{U}}^{ij} Q_{i}^{\text{U}} U_{j}^{\text{U}} H_{2}^{i} + y_{\text{D}}^{ij} Q_{i}^{\text{D}} D_{j}^{\text{D}} H_{1}^{i} + y_{\text{L}}^{ij} L_{i}^{\text{L}} E_{j}^{\text{L}} H_{1}^{i} + \mu H_{1}^{i} H_{2}^{i} \right), \]  

(39)
where $i, j = 1, 2$ are the $SU(2)$ and $a, b = 1, 2, 3$ are the generation indices; the $SU(3)$ colour indices are omitted. This part of the Lagrangian almost exactly repeats that of the SM. The only difference is the last term which describes the Higgs mixing. It is absent in the SM since there is only one Higgs field there.

However, one can write down also the different Yukawa terms

\[ L_{\text{Yukawa}} = \epsilon_{ij} (\lambda_{Labc} L^i_a L^j_b E^c_d + \lambda_{L'abc} L^i_a Q^j_b D^c_d + \mu'_a L^i_a H^2) + \lambda^B_{abc} U^c_a D^b_c D^d_d. \]  

These terms are absent in the SM. The reason is very simple: one can not replace the superfields in Eq. (40) by the ordinary fields like in Eq. (39) because of the Lorentz invariance. These terms have also another property, they violate either the lepton number $L$ (the first 3 terms in Eq. (40)) or the baryon number $B$ (the last term). Since both effects are not observed in Nature, these terms must be suppressed or excluded. One can avoid such terms introducing a new special symmetry called $R$-symmetry [21].

The global $U(1)_R$ invariance

\[ U(1)_R : \theta \rightarrow e^{i \alpha} \theta, \quad \Phi \rightarrow e^{i \alpha} \Phi, \]  

which is reduced to the discrete group $Z_2$, is called $R$-parity. The $R$-parity quantum number is

\[ R = (-1)^{(B-L)+2S} \]  

for the particles with the spin $S$. Thus, all the ordinary particles have the $R$-parity quantum number equal to $R = +1$, while all the superpartners have the $R$-parity quantum number equal to $R = -1$. The first part of the Yukawa Lagrangian is $R$-symmetric, while the second part is $R$-nonsymmetric. The $R$-parity obviously forbids the terms (40). However, it may well be that these terms are present, though experimental limits on the couplings are very severe

\[ \lambda_{Labc}, \quad \lambda_{L'abc} < 10^{-4}, \quad \lambda^B_{abc} < 10^{-9}. \]  

Conservation of the $R$-parity has two important consequences

- the superpartners are created in pairs;
- the lightest superparticle (LSP) is stable. Usually it is the photino $\tilde{\gamma}$, the superpartner of the photon with some admixture of the neutral higgsino. This is the candidate for the DM particle which should be neutral and survive since the Big Bang.

### 4.3 Properties of interactions

If one assumes that the $R$-parity is preserved, then the interactions of the superpartners are essentially the same as in the SM, but two of three particles involved into the interaction at any vertex are replaced by the superpartners. The reason for it is the $R$-parity.

Typical vertices are shown in Fig. 6. The tilde above the letter denotes the corresponding superpartner. Note that the coupling is the same in all the vertices involving the superpartners.

### 4.4 Creation and decay of superpartners

The above-mentioned rule together with the Feynman rules for the SM enables one to draw diagrams describing creation of the superpartners. One of the most promising processes is the $e^+ e^-$ annihilation (see Fig. 7). The usual kinematic restriction is given by the c.m. energy $m_{\text{particle}}^{\text{max}} \leq \sqrt{s}/2$. Similar processes take place at hadron colliders with the electrons and the positrons being replaced by the quarks and the gluons.

Experimental signatures at the hadron colliders are similar to those at the $e^+ e^-$ machines; however, here one has wider possibilities. Besides the usual annihilation channel, one has numerous processes of gluon fusion, quark-antiquark and quark-gluon scattering (see Fig. 8).
Creation of the superpartners can be accompanied by creation of the ordinary particles as well. We consider various experimental signatures below. They crucially depend on the SUSY breaking pattern and on the mass spectrum of the superpartners.

The decay properties of the superpartners also depend on their masses. For the quark and lepton superpartners the main processes are shown in Fig. 9.

5 Breaking of SUSY in the MSSM

Usually it is assumed that the supersymmetry is broken spontaneously via the v.e.v.s of some fields. However, in the case of supersymmetry one can not use the scalar fields like the Higgs field, but rather the auxiliary fields present in any SUSY multiplet. There are two basic mechanisms of spontaneous SUSY breaking: the Fayet-Iliopoulos (or $D$-type) mechanism [22] based on the $D$ auxiliary field from the vector multiplet and the O’Raifeartaigh (or $F$-type) mechanism [23] based on the $F$ auxiliary field from the chiral multiplet. Unfortunately, one can not explicitly use these mechanisms within the MSSM since none of the fields of the MSSM can develop the non-zero v.e.v. without spoiling the gauge invariance. Therefore, the spontaneous SUSY breaking should take place via some other fields.

The most common scenario for producing low-energy supersymmetry breaking is called the hidden sector scenario [24]. According to this scenario, there exist two sectors: the usual matter belongs to the “visible” one, while the second, “hidden” sector, contains the fields which lead to breaking of the
Fig. 8: Examples of diagrams for the SUSY particle production via the strong interactions (top rows for $\tilde{g}\tilde{g}$, $\tilde{q}\tilde{q}$ and $\tilde{g}\tilde{q}$, respectively) and the electroweak interactions (the lowest row).

squarks
$\tilde{q}_{L,R} \rightarrow q + \tilde{\chi}^0_i$
$\tilde{q}_L \rightarrow q' + \tilde{\chi}^0_i$
$\tilde{q}_{L,R} \rightarrow q + \tilde{g}$

sleptons
$\tilde{l} \rightarrow l + \tilde{\chi}^0_i$
$\tilde{l}_L \rightarrow \nu_l + \tilde{\chi}^0_i$

chargino
$\tilde{\chi}^\pm_1 \rightarrow e + \nu_e + \tilde{\chi}^0_i$
$\tilde{\chi}^\pm_1 \rightarrow q + q' + \tilde{\chi}^0_i$

gluino
$\tilde{g} \rightarrow q = q + \tilde{g}$
$\tilde{g} \rightarrow g + \tilde{\gamma}$

neutralino
$\tilde{\chi}^0_1 \rightarrow \tilde{\chi}^0_0 + l^+ + l^-$
$\tilde{\chi}^0_1 \rightarrow \tilde{\chi}^0_0 + q + \tilde{q}'$
$\tilde{\chi}^0_1 \rightarrow \tilde{\chi}^\pm_1 + l^\pm + \nu_l$
$\tilde{\chi}^0_1 \rightarrow \tilde{\chi}^0_1 + \nu_l + \bar{\nu}_l$

Fig. 9: Decay of superpartners
supersymmetry. These two sectors interact with each other by an exchange of some fields called messengers, which mediate SUSY breaking from the hidden to the visible sector. There might be various types of the messenger fields: gravity, gauge, etc. The hidden sector is the weakest part of the MSSM. It contains a lot of ambiguities and leads to uncertainties of the MSSM predictions considered below.

So far there are four known main mechanisms to mediate SUSY breaking from the hidden to the visible sector:

- Gravity mediation (SUGRA) [25];
- Gauge mediation [26];
- Anomaly mediation [27];
- Gaugino mediation [28].

All the four mechanisms of soft SUSY breaking are different in details but are common in results. The predictions for the sparticle spectrum depend on the mechanism of SUSY breaking. In what follows, to calculate the mass spectrum of the superpartners, we need the explicit form of the SUSY breaking terms. For the MSSM without the $R$-parity violation one has in general

$$
\mathcal{L}_{\text{Breaking}} = 
\sum_i m^2_{0i} |\varphi_i|^2 + \left( \frac{1}{2} \sum_{\alpha} M_\alpha \tilde{\lambda}_\alpha \lambda_\alpha + BH_1 H_2 + A_{ab} \tilde{Q}_a \tilde{U}_b H_2 + A_{ab} \tilde{D}_a \tilde{D}_b H_1 + A_\tilde{L}_a \tilde{E}_b \tilde{H}_1 \right),
$$

where we have suppressed the $SU(2)$ indices. Here $\varphi_i$ are all the scalar fields, $\tilde{\lambda}_\alpha$ are the gaugino fields, $\tilde{Q}, \tilde{U}, \tilde{D}$ and $\tilde{L}, \tilde{E}$ are the squark and slepton fields, respectively, and $H_{1,2}$ are the $SU(2)$ doublet Higgs fields.

Eq. (43) contains a vast number of free parameters which spoils the predictive power of the model. To reduce their number, we adopt the so-called universality hypothesis, i.e., we assume the universality or equality of various soft parameters at the high energy scale, namely, we put all the spin-0 particle masses to be equal to the universal value $m_0$, all the spin-1/2 particle (gaugino) masses to be equal to $m_{1/2}$ and all the cubic and quadratic terms, proportional to $A$ and $B$, to repeat the structure of the Yukawa superpotential (39). This is the additional requirement motivated by the supergravity mechanism of SUSY breaking. The universality is not the necessary requirement and one may consider the non-universal soft terms as well. However, it will not change the qualitative picture presented below; so, for simplicity, in what follows we consider the universal boundary conditions. In this case, Eq. (43) takes the form

$$
\mathcal{L}_{\text{Breaking}} = 
m^2_0 \sum_i |\varphi_i|^2 + \left( \frac{m_{1/2}}{2} \sum_{\alpha} \tilde{\lambda}_\alpha \lambda_\alpha + B \mu H_1 H_2 + A \left[ y^{UL}_{ab} \tilde{Q}_a \tilde{U}_b H_2 + y^{DL}_{ab} \tilde{D}_a \tilde{D}_b H_1 + y^{LE}_{ab} \tilde{L}_a \tilde{E}_b \tilde{H}_1 \right] \right).
$$

Thus, we are left with five free parameters, namely, $m_0, m_{1/2}, A, B$ and $\mu$ versus two parameters of the Higgs potential in the SM, $m^2$ and $\lambda$. In the SUSY model the Higgs potential is not arbitrary but is calculated from the Yukawa and gauge terms as we will see below.

The soft terms explicitly break the supersymmetry. As will be shown later, they lead to the mass spectrum of the superpartners different from that of the ordinary particles. Remind that the masses of the quarks and leptons remain zero until the $SU(2)$ symmetry is spontaneously broken.

5.1 The soft terms and the mass formulae

There are two main sources of the mass terms in the Lagrangian: the $D$-terms and the soft ones. With given values of $m_0, m_{1/2}, \mu, Y_t, Y_b, Y_\tau, A,$ and $B$ one can construct the mass matrices for all the particles.
Knowing them at the GUT scale, one can solve the corresponding RG equations, thus linking the values at the GUT and electroweak scales. Substituting these parameters into the mass matrices, one can predict the mass spectrum of the superpartners [29, 30].

5.1.1 Gaugino-higgsino mass terms
The mass matrix for the gauginos, the superpartners of the gauge bosons, and for the higgsinos, the superpartners of the Higgs bosons, is nondiagonal, thus leading to their mixing. The mass terms look like

\[ L_{Gaugino-Higgsino} = -\frac{1}{2} M_3 \tilde{\chi}_a \lambda_a - \frac{1}{2} \tilde{\chi} M^{(0)} \chi - (\tilde{\psi} M^{(c)} \psi + h.c.), \]

where \( \lambda_a, a = 1, 2, \ldots, 8 \) are the Majorana gluino fields and

\[
\chi = \begin{pmatrix} \tilde{B}^0 \\ \tilde{W}^3 \\ \tilde{H}_1^0 \\ \tilde{H}_2^0 \end{pmatrix}, \quad \psi = \begin{pmatrix} \tilde{W}^+ \\ \tilde{H}^+ \end{pmatrix}
\]

are, respectively, the Majorana neutralino and the Dirac chargino fields.

The neutralino mass matrix is

\[
M^{(0)} = \begin{pmatrix} M_1 & 0 & -M_Z \cos \beta \sin \theta_W & M_Z \sin \beta \sin \theta_W \\ 0 & M_2 & M_Z \cos \beta \cos \theta_W & -M_Z \sin \beta \cos \theta_W \\ -M_Z \cos \beta \sin \theta_W & M_Z \cos \beta \cos \theta_W & 0 & -\mu \\ M_Z \sin \beta \sin \theta_W & -M_Z \sin \beta \cos \theta_W & -\mu & 0 \end{pmatrix},
\]

where \( \tan \beta = v_2/v_1 \) is the ratio of two Higgs v.e.v.s and \( \sin \theta_W \) is the usual sine of the weak mixing angle. The physical neutralino masses \( M_{\tilde{\chi}_i} \) are obtained as eigenvalues of this matrix after diagonalization.

For the chargino mass matrix one has

\[
M^{(c)} = \begin{pmatrix} M_2 & \sqrt{2} M_W \sin \beta \\ \sqrt{2} M_W \cos \beta & \mu \end{pmatrix}.
\]

This matrix has two chargino eigenstates \( \tilde{\chi}_i^{\pm} \) with mass eigenvalues

\[
M_{1,2}^2 = \frac{1}{2} \sqrt{M_2^2 + \mu^2 + 2M_W^2 \mp \sqrt{(M_2^2 - \mu^2)^2 + 4M_W^4 \cos^2 2\beta + 4M_W^2(M_2^2 + \mu^2 + 2M_2 \mu \sin 2\beta)}}.
\]

5.1.2 Squark and slepton masses
The non-negligible Yukawa couplings cause mixing between the electroweak eigenstates and the mass eigenstates of the third generation particles. The mixing matrices for \( \tilde{m}_L^2, \tilde{m}_R^2 \) and \( \tilde{m}_\tau^2 \) are

\[
\begin{pmatrix} \tilde{m}_L^2 & m_t (A_t - \mu \cot \beta) \\ m_t (A_t - \mu \cot \beta) & \tilde{m}_R^2 \end{pmatrix},
\]

\[
\begin{pmatrix} \tilde{m}_L^2 & m_b (A_b - \mu \tan \beta) \\ m_b (A_b - \mu \tan \beta) & \tilde{m}_R^2 \end{pmatrix},
\]

\[
\begin{pmatrix} \tilde{m}_L^2 & m_\tau (A_\tau - \mu \tan \beta) \\ m_\tau (A_\tau - \mu \tan \beta) & \tilde{m}_R^2 \end{pmatrix}.
\]
with
\[ \tilde{m}_{1L}^2 = \tilde{m}_{Q}^2 + m_t^2 + \frac{1}{6}(4M_W^2 - M_Z^2) \cos 2\beta, \]
\[ \tilde{m}_{1R}^2 = \tilde{m}_{U}^2 + m_t^2 - \frac{2}{3}(M_W^2 - M_Z^2) \cos 2\beta, \]
\[ \tilde{m}_{bL}^2 = \tilde{m}_{Q}^2 + m_b^2 - \frac{1}{6}(2M_W^2 + M_Z^2) \cos 2\beta, \]
\[ \tilde{m}_{bR}^2 = \tilde{m}_{D}^2 + m_b^2 + \frac{1}{3}(M_W^2 - M_Z^2) \cos 2\beta, \]
\[ \tilde{m}_{1L}^2 = \tilde{m}_{E}^2 + m_t^2 + (M_W^2 - M_Z^2) \cos 2\beta, \]
and the mass eigenstates are the eigenvalues of these mass matrices. For the light generations mixing is negligible.

The first terms here (\(\tilde{m}^2\)) are the soft ones, which are calculated using the RG equations starting from their values at the GUT (Planck) scale. The second ones are the usual masses of the quarks and leptons and the last ones are the \(D\)-terms of the potential.

5.2 The Higgs potential

As has already been mentioned, the Higgs potential in the MSSM is totally defined by the superpotential (and the soft terms). Due to the structure of \(\mathcal{L}_{\text{Yukawa}}\) the Higgs self-interaction is given by the \(D\)-terms while the \(F\)-terms contribute only to the mass matrix. The tree level potential is

\[ V_{\text{tree}} = m_1^2|H_1|^2 + m_2^2|H_2|^2 - m_3^2(H_1H_2 + h.c.) + \frac{g^2 + g'^2}{8}(|H_1|^2 - |H_2|^2)^2 + \frac{g^2}{2}|H_1^*H_2|^2, \]  

(48)

where \(m_1^2 = \mu_1^2, m_2^2 = \mu_2^2 + \mu_3^2\). At the GUT scale \(m_1^2 = m_2^2 = m_0^2 + \mu_0^2\), \(m_3^2 = -B\mu_0\). Notice that the Higgs self-interaction coupling in Eq. (48) is fixed and defined by the gauge interactions as opposed to the Standard Model.

The Higgs scalar potential in accordance with the supersymmetry, is positive definite and stable. It has no nontrivial minimum different from zero. Indeed, let us write the minimization condition for the potential (48)

\[ \frac{1}{2} \frac{\delta V}{\delta H_1} = m_2^2v_1 - m_3^2v_2 + \frac{g^2 + g'^2}{4}(v_1^2 - v_2^2)v_1 = 0, \]
\[ \frac{1}{2} \frac{\delta V}{\delta H_2} = m_3^2v_2 - m_3^2v_1 + \frac{g^2 + g'^2}{4}(v_1^2 - v_2^2)v_2 = 0, \]  

(49)

where we have introduced the notation

\[ \langle H_1 \rangle \equiv v_1 = v \cos \beta, \quad \langle H_2 \rangle \equiv v_2 = v \sin \beta, \]
\[ v^2 = v_1^2 + v_2^2, \quad \tan \beta \equiv \frac{v_2}{v_1}. \]

Solution to Eqs. (49) can be expressed in terms of \(v^2\) and \(\sin 2\beta\)

\[ v^2 = \frac{4(m_1^2 - m_3^2 \tan^2 \beta)}{(g^2 + g'^2)(\tan^2 \beta - 1)}, \quad \sin 2\beta = \frac{2m_3^2}{m_1^2 + m_2^2}. \]  

(50)

One can easily see from Eqn. (50) that if \(m_1^2 = m_2^2 = m_3^2 + \mu_0^2\), \(v^2\) happens to be negative, i. e. the minimum does not exist. In fact, real positive solutions to Eqs. (49) exist only if the following conditions are satisfied:

\[ m_1^2 + m_2^2 > 2m_3^2, \quad m_1^2m_2^2 < m_3^4, \]  

(51)
which is not the case at the GUT scale. This means that spontaneous breaking of the $SU(2)$ gauge invariance, which is needed in the SM to give masses for all the particles, does not take place in the MSSM.

This strong statement is valid, however, only at the GUT scale. Indeed, going down with the energy, the parameters of the potential (48) are renormalized. They become the “running” parameters with the energy scale dependence given by the RG equations.

5.3 Radiative electroweak symmetry breaking

The running of the Higgs masses leads to the remarkable phenomenon known as radiative electroweak symmetry breaking. Indeed, one can see in Fig. 10 that $m_2^2$ (or both $m_1^2$ and $m_2^2$) decreases when going down from the GUT scale to the $M_Z$ scale and can even become negative. As a result, at some value of $Q^2$ the conditions (51) are satisfied, so that the nontrivial minimum appears. This triggers spontaneous breaking of the $SU(2)$ gauge invariance. The vacuum expectations of the Higgs fields acquire nonzero values and provide masses to the quarks, leptons and $SU(2)$ gauge bosons, and additional contributions to the masses of their superpartners.

In this way one also obtains the explanation of why the two scales are so much different. Due to the logarithmic running of the parameters, one needs a long "running time" to get $m_2^2$ (or both $m_1^2$ and $m_2^2$) to be negative when starting from a positive value of the order of $M_{SUSY} \sim 10^2 \div 10^3$ GeV at the GUT scale.

5.4 The superpartners mass spectrum

The mass spectrum is defined by the low energy parameters. To calculate the low energy values of the soft terms, we use the corresponding RG equations [31]. Having all the RG equations, one can now find the RG flow for the soft terms. Taking the initial values of the soft masses at the GUT scale in the interval between $10^2 \div 10^3$ GeV consistent with the SUSY scale suggested by the unification of the gauge couplings (8) leads to the RG flow of the soft terms shown in Fig. 10. [29, 30]

One should mention the following general features common to any choice of initial conditions:

- The gaugino masses follow the running of the gauge couplings and split at low energies. The gluino mass is running faster than the other ones and is usually the heaviest due to the strong interaction.
The squark and slepton masses also split at low energies, the stops (and sbottoms) being the lightest due to the relatively big Yukawa couplings of the third generation.

The Higgs masses (or at least one of them) are running down very quickly and may even become negative.

The typical dependence of the mass spectra on the initial conditions at the GUT scale ($m_0$) is also shown in Fig. 11 [32, 33]. For a given value of $m_{1/2}$ the masses of the lightest particles are practically independent of $m_0$, while the masses of the heavier ones increase with it monotonically. One can see that the lightest neutralinos and charginos as well as the top-squark may be rather light.

5.5 The Higgs boson masses

Provided conditions (51) are satisfied, one can also calculate the masses of the Higgs bosons taking the second derivatives of the potential (48) with respect to the real and imaginary parts of the Higgs fields ($H_i = S_i + iP_i$) in the minimum. The mass matrices at the tree level are

**CP-odd components** $P_1$ and $P_2$:

$$
\mathcal{M}^{\text{odd}} = \left. \frac{\partial^2 V}{\partial P_i \partial P_j} \right|_{H_i = v_i} = \left( \begin{array}{cc} \tan \beta & 1 \\ 1 & \cot \beta \end{array} \right) m_3^2,
$$

(52)

**CP-even neutral components** $S_1$ and $S_2$:

$$
\mathcal{M}^{\text{even}} = \left. \frac{\partial^2 V}{\partial S_i \partial S_j} \right|_{H_i = v_i} = \left( \begin{array}{cc} \tan \beta & -1 \\ -1 & \cot \beta \end{array} \right) m_3^2 + \left( \begin{array}{cc} \cot \beta & -1 \\ -1 & \tan \beta \end{array} \right) m_Z^2 \frac{\sin 2\beta}{2},
$$

(53)

**Charged components** $H^-$ and $H^+$:

$$
\mathcal{M}^{\text{ch}} = \left. \frac{\partial^2 V}{\partial H^- \partial H^+} \right|_{H_i = v_i} = \left( \begin{array}{cc} \tan \beta & 1 \\ 1 & \cot \beta \end{array} \right) \left( m_3^2 + m_W^2 \frac{\sin 2\beta}{2} \right).
$$

(54)
Diagonalizing the mass matrices, one gets the mass eigenstates:

\[
\begin{cases}
G^0 = -\cos \beta P_1 + \sin \beta P_2, & \text{Goldstone boson}\rightarrow Z_0, \\
A = \sin \beta P_1 + \cos \beta P_2, & \text{Neutral CP-odd Higgs},
\end{cases}
\]

\[
\begin{cases}
G^+ = -\cos (\beta H_1^+) + \sin \beta H_2^+, & \text{Goldstone boson}\rightarrow W^+, \\
H^+ = \sin (\beta H_1^+) + \cos \beta H_2^+, & \text{Charged Higgs},
\end{cases}
\]

\[
\begin{cases}
h = -\sin \alpha S_1 + \cos \alpha S_2, & \text{SM CP-even Higgs}, \\
H = \cos \alpha S_1 + \sin \alpha S_2, & \text{Extra heavy Higgs},
\end{cases}
\]

where the mixing angle $\alpha$ is given by

\[
\tan 2\alpha = \tan 2\beta \left( \frac{m_A^2 + M_Z^2}{m_A^2 - M_Z^2} \right).
\]

The physical Higgs bosons acquire the following masses [18]:

**$CP$-odd neutral Higgs $A$:**

\[m_A^2 = m_1^2 + m_2^2, \quad (55)\]

**Charged Higgses $H^\pm$:**

\[m_{H^\pm}^2 = m_A^2 + M_W^2, \quad (56)\]

**$CP$-even neutral Higgses $H, h$:**

\[m_{H,h}^2 = \frac{1}{2} \left[ m_A^2 + M_Z^2 \pm \sqrt{(m_A^2 + M_Z^2)^2 - 4m_A^2 M_Z^2 \cos^2 2\beta} \right], \quad (57)\]

where, as usual,

\[M_W^2 = \frac{g^2 v^2}{2}, \quad M_Z^2 = \frac{g^2 + g'^2}{2} v^2.\]

This leads to the once celebrated SUSY mass relations

\[m_{H^\pm} \geq M_W, \quad m_h \leq m_A \leq M_H, \quad (58)\]

\[m_h \leq M_Z |\cos 2\beta| \leq M_Z, \quad m_h^2 + m_{H^\pm}^2 = m_A^2 + M_Z^2.\]

Thus, the lightest neutral Higgs boson happens to be lighter than the $Z$-boson, which clearly distinguishes it from the SM one. Though we do not know the mass of the Higgs boson in the SM, there are several indirect constraints leading to the lower boundary of $m_h^{SM} \geq 135$ GeV. After including the leading one-loop radiative corrections, the mass of the lightest Higgs boson in the MSSM, $m_h$, reads

\[m_h^2 = M_Z^2 \cos^2 2\beta + \frac{3g^2 m_t^4}{16\pi^2 M_W^2} \log \frac{m_{t_1}^2 m_{t_2}^2}{m_t^4} + \ldots \quad (59)\]

which leads to about 40 GeV increase [34]. The second loop correction is negative but small [35]. It is interesting that the Higgs mass upper bound depends crucially on some parameters of the model, and is almost independent on the choice of the other parameters. For example, the 1 GeV change in the mass of the top quark leads to the ~1 GeV change in the Higgs mass upper bound. The dependence of the maximal Higgs mass on the supersymmetry breaking scale $M_S$ is shown in the left panel of Fig. 12 [36] for different scenarios of SUSY breaking. The widths of bands corresponds to the variation of the top mass in the range 170–176 GeV.

The right panel of Fig. 12 shows the dependence of the maximal Higgs mass on $\tan \beta$ for the fixed value of $m_t = 173$ GeV while other parameters of the model vary within the ranges [37]:

\[\ldots\]
5.6 The lightest superparticle

One of the crucial questions is the properties of the lightest superparticle. Different SUSY breaking scenarios lead to different experimental signatures and different LSP.

- Gravity mediation
  In this case, the LSP is the lightest neutralino $\tilde{\chi}^0_1$, which is almost 90% photino for the low $\tan \beta$ solution and contains more higgsino admixture for high $\tan \beta$. The usual signature for LSP is the missing energy; $\tilde{\chi}^0_1$ is stable and is the best candidate for the cold dark matter particle in the Universe. Typical processes, where the LSP is created, end up with jets + $E_T$, or leptons + $E_T$, or both jets + leptons + $E_T$.

- Gauge mediation
  In this case the LSP is the gravitino $\tilde{G}$, which also leads to the missing energy. The actual question here is what is the NLSP, the next-to-lightest particle, is. There are two possibilities:
  i) $\tilde{\chi}^0_1$ is the NLSP. Then the decay modes are: $\tilde{\chi}^0_1 \rightarrow \gamma \tilde{G}$, $h \tilde{G}$, $Z \tilde{G}$. As a result, one has two hard photons + $E_T$, or jets + $E_T$.
  ii) $\tilde{l}_R$ is the NLSP. Then the decay mode is $\tilde{l}_R \rightarrow \tau \tilde{G}$ and the signature is a charged lepton and the missing energy.

- Anomaly mediation
  In this case, one also has two possibilities:
  i) $\tilde{\chi}^0_1$ is the LSP and wino-like. It is almost degenerate with the NLSP.
  ii) $\tilde{\nu}_L$ is the LSP. Then it appears in the decay of the chargino $\tilde{\chi}^+ \rightarrow \tilde{\nu}_l$ and the signature is the charged lepton and the missing energy.

- R-parity violation
  In this case, the LSP is no longer stable and decays into the SM particles. It may be charged (or even colored) and may lead to rare decays like the neutrinoless double $\beta$-decay, etc.

Experimental limits on the LSP mass follow from the non-observation of the corresponding events. The modern lower limit is around 40 GeV.
6 Constrained MSSM

6.1 Parameter space of the MSSM

The Standard Model has the following set of free parameters:

i) three gauge couplings $\alpha_i$;

ii) three (or four if the Dirac neutrino mass term is included) matrices of the Yukawa couplings $y^i_{ab}$, where $i = U, D, L(N)$;

iii) two parameters of the Higgs potential ($\lambda$ and $m^2$).

The parameters of the Yukawa sector are usually traded for the masses, mixing angles and phases of the mixing matrices.

In the MSSM one has the same set of parameters except for the parameters of the Higgs potential which is fixed by supersymmetry, but in addition one has

iv) the Higgs fields mixing parameter $\mu$;

v) the soft supersymmetry breaking terms.

The main uncertainty comes from the unknown soft terms. With the universality hypothesis one is left with the following set of 5 free parameters defining the mass scales

$$\mu, m_0, m_{1/2}, A \leftrightarrow \tan \beta = \frac{v_2}{v_1}.$$  

When choosing the set of parameters and making predictions, one has two possible ways to proceed:

i) take the low-energy parameters like the superparticle masses $\tilde{m}_{q_1}, \tilde{m}_{q_2}, \tilde{m}_A$, $\tan \beta$, mixings $X_{stop}, \mu$, etc. as input and calculate the cross-sections as functions of these parameters. The disadvantage of this approach is the large number of free parameters.

ii) take the high-energy parameters like the above mentioned 5 parameters as input, run the RG equations and find the low-energy values. Now the calculations can be carried out in terms of the initial parameters. The advantage is that their number is relatively small. A typical range of these parameters is

$$100 \text{ GeV} \leq m_0, m_{1/2}, \mu \leq 3 \text{ TeV},$$

$$-3m_0 \leq A_0 \leq 3m_0, \quad 1 \leq \tan \beta \leq 70.$$  

The experimental constraints are sufficient to determine these parameters, albeit with large uncertainties.

6.2 The choice of constraints

When subjecting constraints on the MSSM, perhaps, the most remarkable fact is that all of them can be fulfilled simultaneously. In our analysis we impose the following constraints on the parameter space of the MSSM:

- LEP II experimental lower limits on the SUSY masses;
- Limits from the Higgs searches;
- Limits from precision measurement of rare decay rates ($B_s \rightarrow \phi \gamma, B_s \rightarrow \mu^+ \mu^-, B_s \rightarrow \tau \nu$);
- Relic abundance of the Dark Matter in the Universe;
- Direct Dark Matter searches;
- Anomalous magnetic moment of the muon;
– Radiative electroweak symmetry breaking;
– Gauge coupling constant unification;
– Neutrality of the LSP;
– Tevatron and LHC limits on the superpartner masses.

In what follows we use the set of experimental data shown in Table 2.

Table 2: List of all constraints used in the fit to determine the excluded region of the CMSSM parameter space.

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Data</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Omega h^2$</td>
<td>$0.113 \pm 0.004$</td>
<td>[38]</td>
</tr>
<tr>
<td>$b \rightarrow s\gamma$</td>
<td>$(3.55 \pm 0.24) \cdot 10^{-4}$</td>
<td>[39]</td>
</tr>
<tr>
<td>$b \rightarrow \tau \nu$</td>
<td>$(1.68 \pm 0.31) \cdot 10^{-4}$</td>
<td>[39]</td>
</tr>
<tr>
<td>$\Delta a_\mu$</td>
<td>$(290 \pm 63(exp) \pm 61(theo)) \cdot 10^{-11}$</td>
<td>[40]</td>
</tr>
<tr>
<td>$B_s \rightarrow \mu \mu$</td>
<td>$B_s \rightarrow \mu \mu &lt; 4.5 \cdot 10^{-9}$</td>
<td>[41]</td>
</tr>
<tr>
<td>$m_h$</td>
<td>$m_h &gt; 114.4$ GeV</td>
<td>[42]</td>
</tr>
<tr>
<td>$m_A$</td>
<td>$m_A &gt; 510$ GeV for $\tan \beta \approx 50$</td>
<td>[43]</td>
</tr>
<tr>
<td>ATLAS</td>
<td>$\sigma^SUSY^ATLAS &lt; 0.001 - 0.03$ pb</td>
<td>[44]</td>
</tr>
<tr>
<td>CMS</td>
<td>$\sigma^SUSY^CMS &lt; 0.003 - 0.03$ pb</td>
<td>[45]</td>
</tr>
<tr>
<td>XENON100</td>
<td>$\sigma^\chi_N &lt; 1.8 \cdot 10^{-45}$ - $6 \cdot 10^{-45}$ cm$^2$</td>
<td>[46]</td>
</tr>
</tbody>
</table>

Having in mind the above mentioned constraints one can find the most probable region of the parameter space by minimizing the $\chi^2$ function [30]. Since the parameter space is 5 dimensional one can not plot it explicitly and is bounded to use various projections. We will accept the following strategy: we first choose the value of the Higgs mixing parameter $\mu$ from the requirement of radiative EW symmetry breaking and then take the plane of parameters $m_0 - m_{1/2}$ adjusting the remained parameters $A_0$ and $\tan \beta$ at each point minimizing the $\chi^2$. We present the restrictions coming from various constraints in the $m_0 - m_{1/2}$ plane.

The most probable region of the parameter space is determined by the minimum $\chi^2_{\text{min}}$ value. The 95% C.L. (90% C.L.) limit is reached for the values of $\chi^2$ of 5.99 (4.61), respectively. The $\chi^2$ function is defined as

$$
\chi^2 = \frac{(\Omega h^2 - 0.1131)^2}{\sigma_{\Omega h^2}^2} + \frac{(b \rightarrow s\gamma - 3.55 \cdot 10^{-4})^2}{\sigma_{b \rightarrow s\gamma}^2} + \frac{(b \rightarrow \tau \nu - 1.68 \cdot 10^{-4})^2}{\sigma_{b \rightarrow \tau \nu}^2} + \frac{(\Delta a_\mu - 302 \cdot 10^{-11})^2}{\sigma_{\Delta a_\mu}^2} + \chi^2_{B_s \rightarrow \mu \mu} + \chi^2_{m_h} + \chi^2_{CMS} + \chi^2_{ATLAS} + \chi^2_{m_A} + \chi^2_{DDMS}
$$

(60)

In what follows we show the influence of various constraints and determine the allowed region of the parameter space with 95% C.L.

7 Electroweak constraints

7.1 Region excluded by the $B_s \rightarrow s\gamma$ decay rate

The next two constraints are related to the rare decays where SUSY may contribute. The first one is the $b \rightarrow s\gamma$ decay which in the SM given by the first two diagrams shown in Fig. 13 and leads to [47]

$$
BR^{SM}(b \rightarrow s\gamma) = (3.15 \pm 0.23) \cdot 10^{-4}
$$

132
while experiment gives \[ \text{BR}^{\text{exp}}(b \to s\gamma) = (3.55 \pm 0.24) \cdot 10^{-4}. \]

These two values almost coincide but still leave some room for SUSY.

SUSY contribution comes from the last three diagrams shown in Fig. 13. The tan $\beta$-enhanced corrections to the chargino and charged Higgs contributions can be summarized as follows: the tan $\beta$-enhanced chargino contributions to BR($b \to s\gamma$) in the large tan $\beta$ regime is [48]

\[
\text{BR}^{\text{SUSY}}(b \to s\gamma) \bigg|_{\tan \beta} \propto \mu_A \tan f(m_{\tilde{\chi}_1^\pm}, m_{\tilde{\chi}_2^\pm}, m_b) \frac{m_b}{v(1 + \delta m_b)},
\]

where all dominant higher-order contributions are included through $\delta m_b$, and $f$ is the integral appearing in the one-loop diagram. The relevant charged-Higgs contributions to $\text{BR}(b \to s\gamma)$ in the large tan $\beta$ regime is [48]

\[
\text{BR}^{\text{SUSY}}(b \to s\gamma) \bigg|_{H^\pm} \propto \frac{m_b(h_t \cos \beta - \Delta h_\tau \sin \beta)}{v \cos \beta(1 + \delta m_b)} g(m_{\tilde{H}_1^\pm}, m_t),
\]

where $g$ is the loop integral appearing in the diagram.

The influence of this constraint is shown below together with the $B_s \to \mu^+\mu^-$ one.

### 7.2 Region excluded by the $B_s \to \mu^+\mu^-$ decay rate

The second example is the $B_s \to \mu^+\mu^-$ decay. In the SM it is given by the first two diagrams shown in Fig. 14. The branching ratio is [49]

\[
\text{BR}^{\text{SM}}(B_s \to \mu^+\mu^-) = (3.23 \pm 0.27) \cdot 10^{-9},
\]

while the recent experiment gives only the lower bound [50] \(^1\)

\[
\text{BR}^{\text{exp}}(B_s \to \mu^+\mu^-) < 4.5 \cdot 10^{-9}.
\]

In the MSSM one has several additional diagrams but the main contribution enhanced by $\tan^6 \beta$ (!) comes from the one shown on the right of Fig. 14.

The branching ratio for $B_s \to \mu^+\mu^-$ is given in [52,53] which we write in the form

\[
\text{BR}(B_s \to \mu^+\mu^-) = \frac{2 \tau_B M_B^2}{64 \pi f_B^2} \sqrt{1 - \frac{4m_f^2}{M_B^2}} \times \left[ \left( 1 - \frac{4m_f^2}{M_B^2} \right) \left( \frac{C_S - C'_S}{m_b + m_s} \right)^2 + \left( \frac{C_P - C'_P}{m_b + m_s} \right)^2 + \frac{2m_f}{M_B^2} \left( C_A - C'_A \right)^2 \right].
\]

\(^1\)While the Lectures have been prepared the first evidence for the decay $B_s \to \mu^+\mu^-$ based on 1.1 fb$^{-1}$ of data recorded in 2012 at $\sqrt{s} = 8$ TeV has been reported [51]. The data show an excess of events with respect to the background-only prediction with a statistical significance of 3.5$\sigma$. A fit to the data gives $\text{BR}(B_s \to \mu^+\mu^-) = (3.2^{+1.5}_{-1.2}) \cdot 10^{-9}$ which is in agreement with the SM prediction, thus leaving less room for SUSY.
Fig. 14: The diagrams contributing to $B_s \to \mu^+ \mu^-$ decay in the SM and in the MSSM

\[
\begin{align*}
B_s & \to \mu^+ \mu^- \\
& \text{SM} \\
& \text{MSSM}
\end{align*}
\]

Fig. 15: The values of the branching ratio of the $B_s \to \mu^+ \mu^-$ decay in the MSSM (left) and constraints on parameters space of the MSSM from electroweak observables (right).

where $f_{B_s}$ is the $B_s$ decay constant, $M_B$ is the $B$-meson mass, $\tau_B$ is the mean life time and $m_l$ is the mass of the lepton. $C_S, C'_S, C_P, C'_P$ include the SUSY loop contributions due to the diagrams involving the particles such as stop, chargino, sneutrino, Higgs etc. For large $\tan \beta$, the dominant contribution to $C_S$ is given approximately by

\[
C_S \simeq \frac{G_F \alpha}{\sqrt{2\pi}} V_{tb} V_{ts}^* \left( \frac{\tan^3 \beta}{4 \sin^2 \theta_W} \right) \left( \frac{m_b m_{\mu} m_{\mu} \mu}{M_W^2 M_A^2} \right) \sin 2\theta_t \left( \frac{m_{t_1}^2 \log \left( \frac{m_{t_1}^2}{\mu^2} \right)}{\mu^2 - m_{t_1}^2} - \frac{m_{t_2}^2 \log \left( \frac{m_{t_2}^2}{\mu^2} \right)}{\mu^2 - m_{t_2}^2} \right)
\]

(64)

where $m_{t_{1,2}}$ are the two stop masses, and $\theta_t$ is the rotation angle to diagonalize the stop mass matrix. We need to multiply the above expression by the factor $1/(1 + \epsilon_b)^2$ to include the SUSY QCD corrections. $\epsilon_b$ is proportional to $\mu \tan \beta$ [54]. Thus, for large $\tan \beta$ the amplitude grows like $\tan^6 \beta$ and might come in contradiction with experiment [55]. One observes, however, that the $\tan \beta$ dependence can be compensated by the strong suppression in the last term if the stop masses become equal. This means that in order to get not too large branching ratio the stop masses have to be degenerate.

The values of the branching ratio for various parameters are shown on the left part of Fig. 15 [53] and the restrictions on the parameter space – on the right of Fig. 15.

7.3 Region excluded by the anomalous magnetic moment of muon

The theoretical value of $g - 2$ has been reviewed in Ref. [56] which is in agreement with the latest values from [57]. Recent measurement of the anomalous magnetic moment of the muon indicates small
deviation from the SM of the order of 3 $\sigma$ [40]:

$$a_\mu^{\text{exp}} = 11 659 2080(63) \cdot 10^{-11}$$
$$a_\mu^{\text{SM}} = 11 659 1790(64) \cdot 10^{-11}$$
$$\Delta a_\mu = a_\mu^{\text{exp}} - a_\mu^{\text{theor}} = (290 \pm 90) \cdot 10^{-11},$$

where the SM contribution comes from

$$a_\mu^{QED} = 11 658 4718.1 (0.2) \cdot 10^{-11}$$
$$a_\mu^{\text{weak}} = 153.2 (1.8) \cdot 10^{-11}$$
$$a_\mu^{\text{hadron}} = 6918.7 (65) \cdot 10^{-11},$$

so that the accuracy of the experiment finally reaches the order of the weak contribution. The corresponding diagrams are shown in Fig. 16.

The deficiency may be easily filled with the SUSY contribution coming from the last two diagrams of Fig. 16. They are similar to that of the weak interactions after replacing the vector bosons by the corresponding diagrams are shown in Fig. 16.

The total contribution to $a_\mu$ from these diagrams is [58]

$$a_\mu^{\text{SUSY}} = -\frac{g_2^2}{8\pi^2} \left\{ \sum_{\chi_i^0, \tilde{\chi}_j^0} \frac{m_{\mu}}{m_{\chi_i^0}} \left[ (-1)^j+1 \sin 2\theta B_1(\eta_{ij}) \tan \theta W N_{i1} \left[ \tan \theta W N_{i1} + N_{i2} \right] \right] \\
+ \frac{m_{\mu}}{2M_W \cos \beta} B_1(\eta_{ij})N_{i3} \left[ 3 \tan \theta W N_{i1} + N_{i2} \right] \\
+ \left( \frac{m_{\mu}}{m_{\chi_i^0}} \right)^2 A_1(\eta_{ij}) \left\{ \frac{1}{4} \left[ \tan \theta W N_{i1} + N_{i2} \right]^2 + \left[ \tan \theta W N_{i1} \right]^2 \right\} \right\} \right\} \tag{65}$$

where $N_{ij}$ are elements of the matrix diagonalizing the neutralino mass matrix, and $U_{ij}, V_{ij}$ are the corresponding ones for the chargino mass matrix, the functions $A$ and $B$ are the one-loop triangle integrals.

For large $\tan \beta$ it can be approximated as [59]

$$|a_\mu^{\text{SUSY}}| \simeq \frac{\alpha(M_Z)}{8\pi \sin^2 \theta_W} \frac{m_{\mu}^2 \tan \beta}{m_{\text{SUSY}}^2} \left( 1 - \frac{4\alpha}{\pi} \log \frac{m_{\text{SUSY}}}{m_{\mu}} \right) \approx 14.0 \cdot 10^{-10} \left( \frac{100 \text{ GeV}}{m_{\text{SUSY}}} \right)^2 \tan \beta,$$

where $m_{\mu}$ is the muon mass, $m_{\text{SUSY}}$ is an average mass of the supersymmetric particles in the loop (essentially the chargino mass). It is proportional to $\mu$ and $\tan \beta$ and requires the positive sign of $\mu$ that kills a half of the parameter space of the MSSM [60].
If the SUSY particles are light enough they give the desired contribution to the anomalous magnetic moment. However, if they are too light the contribution exceeds the gap between the experiment and the SM. For too heavy particles the contribution is too small. The values of $a_{\mu}^{SUSY}$ versus $\tan \beta$ for various values of the SUSY breaking parameters $m_0$ and $m_{1/2}$ are shown on the left of Fig. 17 and the restrictions on the parameter space are presented on the right panel of Fig. 17. However, the allowed region is almost excluded by the direct SUSY searches at the LHC as can be seen in Fig. 17 on the right panel. So the observed deviation from the SM might be caused by the other reasons.

### 7.4 Region excluded by the pseudo-scalar Higgs mass $m_A$

The pseudo-scalar Higgs boson production is enhanced by $\tan \beta$. The main diagrams for for the gluon fusion and associated Higgs production with a $b$-quark are shown in Fig. 18 together with the corresponding cross-sections. Since the $b$-quark production is mostly in the forward direction, the scale on the right-hand side indicates if at least one $b$-quark is required to be in the acceptance, defined by $\eta < 2.5$, and have a transverse momentum above 20 GeV/c.

The pseudo-scalar Higgs boson mass is determined by the relic density constraint, because the dominant neutralino annihilation contribution comes from the $A$-boson exchange in the region outside the small co-annihilation regions. One expects $m_A \propto m_{1/2}$ from the relic density constraint, which can be fulfilled with $\tan \beta$ values around 50 in the whole $(m_0 - m_{1/2})$ plane [61]. Since the $A$ production cross section at the LHC is proportional to $\tan^2 \beta$ the pseudo-scalar mass limit increases up to 496 GeV for the large values of $\tan \beta$ preferred by the relic density (see Fig. 18 right panel). The corresponding $m_A$-values are displayed in the left panel of Fig. 19 and the $m_A$ values excluded by the LHC searches lead to the excluded region, shown by the contour line in Fig. 19.

The rather strong limits on the pseudo-scalar Higgs boson mass from LHC [43, 65], especially at large values of $\tan \beta$, lead then to constraints on $m_{1/2}$ of about 400 GeV for intermediate values of $m_0$, as shown in the left panel of Fig. 19 [66].

### 7.5 Effect of a SM Higgs mass $m_h$ around 125 GeV

The 95% C.L. LEP limit of 114.4 GeV contributes for the small and intermediate SUSY masses to the $\chi^2$ function. In recent publications CMS [67] and ATLAS [68] collaborations show evidence for the
Higgs with a mass around 125 GeV. If we assume this to be the evidence for the SM Higgs boson, which has similar properties as the lightest SUSY Higgs boson in the decoupling regime, we can check the consequences for the CMSSM. If the Higgs mass of 125 GeV is included to the fit, the best-fit point moves to higher SUSY masses, but there is rather strong tension between the relic density constraint, $B_s \to \mu^+\mu^-$ and the Higgs mass, so the best-fit point depends strongly on the error assigned to the Higgs mass, as shown in Fig. 19 (right panel). The experimental error on the Higgs mass is about 2 GeV, but the theoretical error can be easily 3 GeV. Therefore, we have plotted the best-fit point for Higgs uncertainties between 2 and 5 GeV. One sees that the best-fit points wanders by several TeV. Clearly this needs a more detailed investigation in the future. It should be noted that the fit does not provide the maximum mixing scenario. If we exclude all other constraints, the maximum value of the Higgs mass can reach 125 GeV, albeit also at similarly large values of $m_{1/2}$. A negative sign of the mixing parameter $\mu$ shows similar results [66].

8 The problem of the dark matter in the Universe

As has been already mentioned the shining matter does not compose all the matter in the Universe. According to the latest precise data [69] the matter content of the Universe is the following:

$$\Omega_{_{total}} = 1.02 \pm 0.02,$$
$$\Omega_{_{vacuum}} = 0.73 \pm 0.04,$$
$$\Omega_{_{matter}} = 0.23 \pm 0.04,$$
$$\Omega_{_{baryon}} = 0.044 \pm 0.004,$$

so that the dark matter prevails the usual baryonic matter by factor of 6.

Besides the rotation curves of spiral galaxies the dark matter manifests itself in the observation of gravitational lensing effects [11] and the large structure formation. It is believed that the dark matter played the crucial role in the formation of large structures like clusters of galaxies and the usual matter just fell down in a potential well attracted by the gravitational interaction afterwards. The dark matter can not make compact objects like the usual matter since it does not take part in the strong interaction.
Fig. 19: Left: values of $m_A$ in the $(m_0 - m_1/2)$ plane after optimizing $\tan \beta$ and $A_0$. The region below the solid line is excluded at 95% C.L. Right: The influence of a Higgs mass of 125 GeV. If it is imposed in the fit, the best-fit point moves to higher SUSY masses, but the location is strongly dependent on the assumed error for the calculated Higgs mass. This error is indicated by the number inside the circle for the best-fit point: $\Delta \chi^2 = 5.99(2\sigma)$ contour.

and can not lose energy by the photon emission since it is neutral. For this reason the dark matter can be trapped in much larger scale structures like galaxies.

In general one may assume two possibilities: either the dark matter interacts only gravitationally, or it participates also in the weak interaction. The latter case is preferable since then one may hope to detect it via the methods of the particle physics. What makes us to believe that the dark matter is probably the Weakly Interacting Massive Particle (WIMP)? This is because the cross-section of the DM annihilation which can be figured out of the amount of the DM in the Universe is close to a typical weak interaction cross-section. Indeed, let us assume that all the DM is made of particles of a single type. Then the amount of the DM can be calculated from the Boltzmann equation [70, 71]

$$\frac{dn_\chi}{dt} + 3Hn_\chi = -\langle \sigma v \rangle (n_\chi^2 - n_{\chi,eq}^2),$$

(66)

where $H = \dot{R}/R$ is the Hubble constant and $n_{\chi,eq}$ is the equilibrium concentration. The relic abundance is expressed in terms of $n_\chi$ as

$$\Omega_\chi h^2 = \frac{m_\chi n_\chi}{\rho_c} \approx \frac{2 \cdot 10^{-27} \text{ cm}^3 \text{ sec}^{-1}}{\langle \sigma v \rangle}.$$  

(67)

Having in mind that $\Omega_\chi h^2 \approx 0.113 \pm 0.009$ and $v \sim 300 \text{ km/sec}$ one gets

$$\sigma \approx 10^{-34} \text{ cm}^2 = 100 \text{ pb},$$

(68)

which is a typical electroweak cross-section.

8.1 Supersymmetric interpretation of the Dark Matter

Supersymmetry offers several candidates for the role of the cold dark matter particle. If one looks at the particle content of the MSSM from the point of view of a heavy neutral particle, one finds several such particles, namely: the superpartner of the photon (the photino $\tilde{\gamma}$), the superpartner of the $Z$-boson (the particle called zino $\tilde{z}$), the superpartner of the neutrino (the sneutrino $\tilde{\nu}$) and the superpartners of the Higgs bosons (the higgsinos $\tilde{H}$). The DM particle can be the lightest of them, the LSP. The others decay
to the LSP and the SM particles, while the LSP is stable and can survive since the Big Bang. As a rule the lightest supersymmetric particle is the neutralino, the spin 1/2 particle which is the combination of the photino, zino and two neutral higgsinos and is the eigenstate of the mass matrix

$$|\chi_1^0\rangle = N_1 |B_0\rangle + N_2 |W_0^3\rangle + N_3 |H_1\rangle + N_4 |H_2\rangle.$$ 

Thus, supersymmetry actually predicts the existence of the dark matter. Moreover, one can choose the parameters of soft supersymmetry breaking in such a way that one gets the right amount of the DM. This requirement serves as a constraint for these parameters and is consistent with the requirements coming from the particle physics.

The search for the LSP was one of the tasks of LEP. They were supposed to be produced as a result of the chargino decays and be detected via the missing transverse energy and momentum. Negative results defined the limit on the LSP mass as shown in Fig. 20.

The DM particles which form the halo of the galaxy annihilate to produce the ordinary particles in the cosmic rays. Identifying them with the LSP from a supersymmetric model one can calculate the annihilation rate and study the secondary particle spectrum. The dominant annihilation diagrams of the neutralino LSP are shown in Fig. 21. The usual final states are either the quark-antiquark pairs or the W and Z bosons. Since the cross sections are proportional to the final state fermion mass, the heavy fermion final states, i.e., the third generation quarks and leptons, are expected to be dominant. The W and Z final states from the t-channel chargino and neutralino exchange have usually a smaller cross section.

The dominant contribution comes from the A-boson exchange: $\chi + \chi \to A \to b\bar{b}$. The sum of the diagrams should yield $\langle\sigma v\rangle = 2 \cdot 10^{-26}$ cm$^3$/sec to get the correct relic density.

The spectral shape of the secondary particles the from DM annihilation is well known from the fragmentation of the mono-energetic quarks studied at the electron-positron colliders, like LEP at CERN, which has been operating up to the centre-of-mass energy of about 200 GeV, i.e. it corresponds to the neutralino mass up to 100 GeV. The different quark flavours all yield similar gamma spectra at high energies. Hence, the spectra of the positrons, photons and antiprotons is known. The relative amount of $\gamma$, $p^+$ and $e^+$ is also known. One expects around 37 photons per collision.

The gamma rays from the DM annihilation can be distinguished from the background by their completely different spectral shape: the background originates mainly from cosmic rays hitting the gas of the disc and producing abundantly $\pi^0$-mesons, which decay into two photons. The initial cosmic ray spectrum is a steep power law spectrum, which yields a much softer gamma ray spectrum than the
fragmentation of the hard mono-energetic quarks from the DM annihilation. The spectral shape of the gamma rays from the background is well known from fixed target experiments given the known cosmic ray spectrum.

Unfortunately, modern data on diffuse galactic gamma rays, do not indicate statistically significant departure from the background. Local excess observed in some experiments like EGRET space telescope [73] and FERMI [74] is well inside the uncertainties of the background.

8.2 Region excluded by the relic density

The observed relic density of the dark matter corresponds to \( \Omega h^2 = 0.113 \pm 0.004 \) [38]. This number is inversely proportional to the annihilation cross section. The dominant annihilation contribution comes from \( A \)-boson exchange in most of the parameter space. The cross section for \( \chi + \chi \rightarrow A \rightarrow b\bar{b} \) can be written as:

\[
\langle \sigma v \rangle \sim \frac{M_1^4 m_A^2 \tan^2 \beta (N_{31} \sin \beta - N_{41} \cos \beta)^2 (N_{21} \cos \theta_W - N_{11} \sin \theta_W)^2}{\sin^2 \theta_W M_Z^2 (4M_1^2 - M_A^2)^2 + M_A^2 \Gamma_A^2}. \tag{69}
\]

As have been mentioned, the correct relic density requires \( \langle \sigma v \rangle = 2 \cdot 10^{-26} \) cm\(^3\)/s, which implies that the annihilation cross section \( \sigma \) is of the order of a 100 pb. Such a high cross section can be obtained only close to the resonance. Actually on the resonance the cross section is too high, so one needs to be in the tail of the resonance, i.e., \( m_A \approx 2.2 m_\chi \) or \( m_A \approx 1.8 m_\chi \). So one expects \( m_A \propto m_{1/2} \) from the relic density constraint. This constraint can be fulfilled with \( \tan \beta \) values around 50 in the whole \((m_0 - m_{1/2})\) plane, except for the narrow co-annihilation regions [61]. The results can be extended to larger values of \( m_0 \), as shown in the left panel of Fig. 22 [75].

8.3 Region excluded by direct DM searches

There are two methods of the dark matter detection: direct and indirect. In the direct detection one assumes that the particles of the dark matter to the Earth and interact with the nuclei of a target. In the underground experiments one can hope to observe such events measuring the recoil energy. There are several experiments of this type: DAMA, Zeplin, CDMS and Edelweiss. Among them only the DAMA collaboration claims to observe a positive outcome in the annual modulation of the signal with the fitted dark matter particle mass around 50 GeV [76].

All the other experiments do not see it though CDMS collaboration recently announced about a few events of a desired type [77]. The reason of this disagreement might be in the different methodology and the targets used since the cross-section depends on the spin of the target nucleus. The collected statistics is also essentially different. DAMA has accumulated by far more data and this is the only experiment which studies the modulation of the signal that may be crucial for reducing the background.
Fig. 22: Left: Fitted values of $\tan \beta$ in the $(m_0 - m_{1/2})$ plane after optimizing $A_0$ to fulfil the relic density and EWSB constraints at every point. The relic density requires $\tan \beta \approx 50$ in most of the parameter space, where pseudo-scalar Higgs exchange dominates. In the (non-red) edges where $\tan \beta$ is lower, the co-annihilation contributes. Right: $\Delta \chi^2 = \chi^2 - \chi^2_{\text{min}}$ distribution in the $(m_0 - m_{1/2})$ plane after imposing the electroweak constraints in comparison with the XENON100 limits [46] on the direct WIMP-nucleon cross-section for the two values of the form factors (dotted line: $\pi N$ scattering, dashed dotted line: lattice gauge theories).

Fig. 23: The exclusion plots from the direct dark matter detection experiments. The spin-independent case (left) from Chicagoland Observatory for Underground Particle Physics (COUPP) and the spin-dependent case (right) from Cryogenic Dark Matter Search (CDMS).

The cross section for direct scattering of WIMPS on nuclei has an experimental upper limit of about $10^{-8}$ pb, i.e. many orders of magnitude below the annihilation cross section. This cross section is related to the annihilation cross section by similar Feynman diagrams. The many orders of magnitude are naturally explained in Supersymmetry by the fact that both cross sections are dominated by Higgs exchange and the fact that the Yukawa couplings to the valence quarks in the proton or neutron are negligible. Most of the scattering cross section comes from the heavier sea-quarks. However, the density of these virtual quarks inside the nuclei is small, which is one of the reasons for the small elastic scattering cross section. In addition, the momentum transfer in elastic scattering is small, so the propagator leads to a cross section inversely proportional to the fourth power of the Higgs mass.

The typical exclusion plots for the spin-independent and spin-dependent cross-sections are shown in Fig. 23 where one can see DAMA allowed region overlapping with the other exclusion curves. Still today we have no convincing evidence for direct dark matter detection or exclusion. Scattering of the
LSP on nuclei can only happen via elastic scattering, provided \( R \)-parity is conserved [13, 70]. The corresponding diagrams are shown in Fig. 24.

The big blob indicates that one enters a low energy regime, in which case the protons and neutrons inside the nucleus cannot be resolved. In this case the spin-independent scattering becomes coherent on all nuclei and the cross section becomes proportional to the number of nuclei:

\[
\sigma = \frac{4}{\pi} \frac{m_{\text{DM}}^2 m_N^2}{(m_{\text{DM}} + m_N)^2} \left( Z f_p + (A - Z) f_n \right)^2
\]  

(70)

where \( A \) and \( Z \) are the atomic mass and atomic number of the target nuclei and the form factors are [78]

\[
f_{p,n} = \sum_{q=u,d,s} G_q f_{Tq}^{(p,n)} \frac{m_{p,n}}{m_q} + \frac{2}{27} f_{TG}^{(p,n)} \sum_{q=c,b,t} G_q \frac{m_{p,n}}{m_q},
\]

(71)

where \( G_q = \lambda_{\text{DM}} \lambda_q / M_f^2 \). Here \( M \) denotes the mediator, and \( \lambda_{\text{DM}}, \lambda_f \) denote the mediator’s couplings to DM and quark. The parameters \( f_{Tq}^{(p)} \) are defined by

\[
m_p f_{Tq}^{(p)} = \langle p | m_q \bar{q} q | p \rangle
\]

(72)

and similar for \( f_{Tq}^{(n)} \). Since the particle which mediates the scattering is typically much heavier than the momentum transfer, the scattering can be written in terms of an effective coupling, which can be determined phenomenologically from \( \pi N \) scattering or from lattice QCD calculations.

The default values of the effective couplings in micrOMEGAs [79] are: \( f_{Tq}^{(p,n)} = 0.033, f_{Tq}^{(p)} = 0.023, f_{Tq}^{(n)} = 0.042, f_{Tq}^{(n)} = 0.018, f_{Tq}^{(n)} = 0.26 \). The lower values from the lattice calculations [80] are: \( f_{Tq}^{(p,n)} = 0.020, f_{Tq}^{(p)} = 0.026, f_{Tq}^{(n)} = 0.02, f_{Tq}^{(n)} = 0.014, f_{Tq}^{(n)} = 0.036, f_{Tq}^{(n)} = 0.02 \). Hence the most important coupling to the strange quarks vary from 0.26 to 0.02 [81], which implies an order of magnitude uncertainty in the elastic neutralino-nucleon scattering cross section.

Another normalization uncertainty in direct dark matter experiments arises from the uncertainty in the local DM density, which can take values between 0.3 and 1.3 GeV/cm\(^3\), as determined from the rotation curve of the Milky Way, see Ref. [82–85].

The excluded region from the XENON100 cross section limit [46] for two choices of form factors is shown in Fig. 22. At large values of \( m_0 \) EWSB forces the higgsino component of the WIMP to increase and consequently the exchange via the Higgs, which has an amplitude proportional to the bino-higgsino mixing, starts to increase. This leads to an increase in the excluded region at large \( m_0 \) and has here a similar sensitivity as the LHC. If we would take the less conservative effective couplings from the default values of micrOMEGAs the XENON100 limit would be 50% higher than the LHC limit [66].
9 Search for SUSY at colliders

9.1 Experimental signatures at $e^+e^-$ colliders

Experiments are finally beginning to push into a significant region of supersymmetry parameter space. We know the sparticles and their couplings, but we do not know their masses and mixings. Given the mass spectrum one can calculate the cross-sections and consider the possibilities of observing the new particles at modern accelerators. Otherwise, one can get restrictions on the unknown parameters.

We start with the $e^+e^-$ colliders. In the leading order the processes of creation of the superpartners are given by the diagrams shown in Fig. 7 above. For a given center of mass energy the cross-sections depend on the masses of the created particles and vanish at the kinematic boundary. Experimental signatures are defined by the decay modes which vary with the mass spectrum. The main ones are summarized below, see, e. g. [19, 86]

<table>
<thead>
<tr>
<th>Production</th>
<th>Decay Modes</th>
<th>Signatures</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tilde{\ell}<em>{L,R}\tilde{\ell}</em>{L,R}$</td>
<td>$\tilde{\ell}_R^\pm \rightarrow l^\pm \tilde{\chi}_1^0$</td>
<td>acompl pair of</td>
</tr>
<tr>
<td></td>
<td>$\tilde{\ell}_L^\pm \rightarrow l^\pm \tilde{\chi}_1^0$</td>
<td>charged lept + $\not{E}_T$</td>
</tr>
<tr>
<td>$\tilde{\nu}\tilde{\nu}$</td>
<td>$\tilde{\nu} \rightarrow l^\pm \tilde{\chi}_1^0$</td>
<td>$\not{E}_T$</td>
</tr>
<tr>
<td>$\tilde{\chi}_1^\pm \tilde{\chi}_1^\mp$</td>
<td>$\tilde{\chi}_1^\pm \rightarrow \tilde{\chi}_1^0 l^\pm \nu$</td>
<td>iso lept + 2 jets + $\not{E}_T$</td>
</tr>
<tr>
<td></td>
<td>$\tilde{\chi}_1^\pm \rightarrow \tilde{\chi}_1^0 f f'$</td>
<td>pair of acompl</td>
</tr>
<tr>
<td></td>
<td>$\rightarrow l\nu\tilde{\chi}_1^0$</td>
<td>leptons + $\not{E}_T$</td>
</tr>
<tr>
<td></td>
<td>$\rightarrow \nu l\tilde{\chi}_1^0$</td>
<td>4 jets + $\not{E}_T$</td>
</tr>
<tr>
<td>$\tilde{\chi}_1^0 \tilde{\chi}_j^0$</td>
<td>$\tilde{\chi}_1^0 \rightarrow \tilde{\chi}_1^0 X$</td>
<td>$X = \nu l\tilde{\chi}_1^0$ invisible</td>
</tr>
<tr>
<td></td>
<td>$= \gamma, 2l, 2$ jets</td>
<td>$\rightarrow 2l + \not{E}_T, l + 2j + \not{E}_T$</td>
</tr>
<tr>
<td>$\tilde{t}\tilde{t}$</td>
<td>$\tilde{t}_1 \rightarrow c\tilde{\chi}_1^0$</td>
<td>2 jets + $\not{E}_T$</td>
</tr>
<tr>
<td></td>
<td>$\tilde{t}_1 \rightarrow b\tilde{\chi}_1^\pm$</td>
<td>2 b-jets + 2 lept + $\not{E}_T$</td>
</tr>
<tr>
<td></td>
<td>$\rightarrow bf\tilde{\chi}_1^0$</td>
<td></td>
</tr>
<tr>
<td>$\tilde{b}\tilde{b}$</td>
<td>$\tilde{b}_i \rightarrow b\tilde{\chi}_1^0$</td>
<td>2 b-jets + $\not{E}_T$</td>
</tr>
<tr>
<td></td>
<td>$\tilde{b}_i \rightarrow b\tilde{\chi}_2^0$</td>
<td>2 b-jets + 2 lept + $\not{E}_T$</td>
</tr>
<tr>
<td></td>
<td>$\rightarrow bf\tilde{\chi}_1^0$</td>
<td>2 b-jets + 2 jets + $\not{E}_T$</td>
</tr>
</tbody>
</table>

The characteristic feature of all the possible signatures is the missing energy and transverse momentum, which is a trade mark of the new physics.

Numerous attempts to find the superpartners at LEP II gave no positive result thus imposing the lower bounds on their masses [87]. Typical LEP II limits on the superpartner masses are

$$m_{\chi_1^0} > 40 \text{ GeV}, \ m_{\tilde{e}} > 105 \text{ GeV}, \ m_{\tilde{\tau}} > 90 \text{ GeV}$$
$$m_{\chi_1^\pm} > 100 \text{ GeV}, \ m_{\tilde{\mu}} > 100 \text{ GeV}, \ m_{\tilde{\tau}} > 80 \text{ GeV}, \ m_{\tilde{\tau}} > 80 \text{ GeV}$$


9.2 Experimental signatures at hadron colliders

Experimental SUSY signatures at the Tevatron and LHC are similar. The strategy of the SUSY searches is based on the assumption that the masses of the superpartners indeed are in the region of 1 TeV so that they might be created on the mass shell with the cross-section big enough to distinguish them from the background of the ordinary particles. Calculation of the background in the framework of the Standard Model thus becomes essential since the secondary particles in all the cases are the same.

There are many possibilities to create the superpartners at the hadron colliders. Besides the usual annihilation channel there are numerous processes of the gluon fusion, quark-antiquark and quark-gluon scattering. The maximal cross-sections of the order of a few picobarn can be achieved in the process of gluon fusion.

As a rule all the superpartners are short lived and decay into the ordinary particles and the lightest superparticle. The main decay modes of the superpartners which serve as the manifestation of SUSY are

<table>
<thead>
<tr>
<th>Production</th>
<th>Decay Modes</th>
<th>Signatures</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g g, q q, g q$</td>
<td>$g \rightarrow q q \tilde{\chi}^0_1$</td>
<td>$E_T$ + multijets (+leptons)</td>
</tr>
<tr>
<td>$\tilde{\chi}^{\pm}_1 \tilde{\chi}^0_2$</td>
<td>$\tilde{\chi}^{\pm}_1 \rightarrow \tilde{\chi}^0_1 l^\pm l$</td>
<td>Trilepton + $E_T$</td>
</tr>
<tr>
<td>$\tilde{\chi}^+_1 \tilde{\chi}^-_1$</td>
<td>$\tilde{\chi}^+_1 \rightarrow l \tilde{\chi}^0_1 l^\pm l$</td>
<td>Dilepton + jet + $E_T$</td>
</tr>
<tr>
<td>$\tilde{\chi}^0_1 \tilde{\chi}^-_1$</td>
<td>$\tilde{\chi}^-_1 \rightarrow \tilde{\chi}^0_1 q q$</td>
<td>Dilepton + $E_T$</td>
</tr>
<tr>
<td>$\tilde{\chi}^0_1 \tilde{\chi}^0_1$</td>
<td>$\tilde{\chi}^0_1 \rightarrow \tilde{\chi}^0_1 X$</td>
<td>$E_T$ + Dilept+jets</td>
</tr>
<tr>
<td>$t \bar{t}$</td>
<td>$t \rightarrow c \tilde{\chi}^0_1$</td>
<td>2 acollin jets + $E_T$</td>
</tr>
<tr>
<td>$t \bar{t}$</td>
<td>$t \rightarrow b \tilde{\chi}^+_1$</td>
<td>Sing lept + $E_T$ + $b$'s</td>
</tr>
<tr>
<td>$t \bar{t}$</td>
<td>$t \rightarrow b \tilde{\chi}^+ l q$</td>
<td>Dilept + $E_T$ + $b$'s</td>
</tr>
<tr>
<td>$t \bar{t}$</td>
<td>$t \rightarrow b \tilde{\chi}^+ l q$</td>
<td>Dilept + $E_T$ + $b$'s</td>
</tr>
<tr>
<td>$l \bar{l}, \bar{l} \nu, \bar{\nu} \nu$</td>
<td>$l \rightarrow l \pm \tilde{\chi}^0_1$</td>
<td>Dilepton + $E_T$</td>
</tr>
<tr>
<td>$l \bar{l}$</td>
<td>$l \rightarrow l \pm \tilde{\chi}^0_1$</td>
<td>Single lepton + $E_T$</td>
</tr>
<tr>
<td>$\bar{\nu}$</td>
<td>$\bar{\nu} \rightarrow \nu \tilde{\chi}^0_1$</td>
<td>$E_T$</td>
</tr>
</tbody>
</table>

Note again the typical events with the missing energy and transverse momentum that is the main difference from the background processes of the Standard Model. Contrary to the $e^+ e^-$ colliders, at hadron machines the background is extremely rich and essential. The missing energy is carried away by the heavy particle with the mass of the order of 100 GeV that is essentially different from the processes with the neutrino in the final state. In hadron collisions the superpartners are always created in pairs and then further quickly decay creating a cascade with the ordinary quarks (i.e. hadron jets) or leptons in the final state plus the missing energy. For the case of the gluon fusion with the creation of gluino it is presented in Table 3 (right panel).

The chargino and neutralino can also be produced in pairs through the Drell-Yang mechanism $pp \rightarrow \tilde{\chi}^{\pm}_1 \tilde{\chi}^0_1$ and can be detected via their lepton decays $\tilde{\chi}^{\pm}_1 \tilde{\chi}^0_1 \rightarrow l l l$ + $E_T$. Hence the main signal
of their creation is the isolated leptons and the missing energy, see Table 3 (left panel). The main background in the trilepton channel comes from the creation of the standard particles $WZ/ZZ, tt, Zb\bar{b}$. There might be also the supersymmetric background from the cascade decays of the squarks and gluinos in multilepton modes.

9.3 Excluded region by direct searches for SUSY at the LHC

The background from the SM processes results in the same final states although with different kinematics. The missing energy in this case is taken away by the light neutrinos. The corresponding processes are shown in Table 4.

Numerous SUSY searches have been already performed at the Tevatron. The pair-produced squarks and gluinos have at least two large-$E_T$ jets associated with the large missing energy. The final state with the lepton(s) is possible due to the leptonic decays of the $\tilde{\chi}^\pm_1$ and/or $\tilde{\chi}^0_2$.

In the trilepton channel the Tevatron is sensitive up to $m_{1/2} \leq 250$ GeV if $m_0 \leq 200$ GeV and up to $m_{1/2} \leq 200$ GeV if $m_0 \geq 500$ GeV. The existing limits on the squark and gluino masses at the Tevatron are [88]: $m_{\tilde{q}} \geq 300$ GeV, $m_{\tilde{g}} \geq 195$ GeV.

In the proton-proton collisions at the LHC the supersymmetric particles can be produced according to the main diagrams shown in the first three rows of Fig. 8, while the main diagrams for the electroweak production are shown in the last row. The corresponding cross-sections are shown in Fig. 25 for the centre-of-mass energy of 7 TeV [75]. One observes that the cross-section for the “strong” production of $\tilde{q}\tilde{q}$ and $\tilde{g}\tilde{q}$ are large for the low values of $m_0$ and $m_{1/2}$, the gluino production $\tilde{g}\tilde{g}$ is the strongest at the small values of $m_{1/2}$ and the electroweak production of gauginos starts to increase at the large values of $m_0$. The reason for the increase of the electroweak production at large $m_0$ is the decrease of the Higgs mixing parameter $\mu$, as determined from the EWSB, which leads to stronger mixing of the Higgsino.
Table 3: Creation of the lightest chargino and the second neutralino with further cascade decay (left). Creation of the pair of gluinos with further cascade decay (right).

<table>
<thead>
<tr>
<th>Process</th>
<th>Final states</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p(q))</td>
<td>(2\ell) (2\nu) (\not{E}_T)</td>
</tr>
<tr>
<td>(\tilde{p}(\tilde{q}))</td>
<td>(\ell) (\nu) (\not{E}_T)</td>
</tr>
<tr>
<td>(p(\tilde{q}))</td>
<td>(3\ell) (\nu) (\not{E}_T)</td>
</tr>
<tr>
<td>(p(\tilde{q}))</td>
<td>(\ell) (2j) (\not{E}_T)</td>
</tr>
<tr>
<td>(p(\tilde{q}))</td>
<td>(\ell) (\nu) (\not{E}_T)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Process</th>
<th>Final states</th>
</tr>
</thead>
<tbody>
<tr>
<td>(g) (\tilde{g}) (\tilde{b})</td>
<td>(2\ell) (2\nu) (6j) (\not{E}_T)</td>
</tr>
<tr>
<td>(g) (\tilde{g}) (\tilde{b})</td>
<td>(2\ell) (6j) (\not{E}_T)</td>
</tr>
<tr>
<td>(g) (\tilde{g}) (\tilde{b})</td>
<td>(2\ell) (8j) (\not{E}_T)</td>
</tr>
<tr>
<td>(g) (\tilde{g}) (\tilde{b})</td>
<td>(8j) (\not{E}_T)</td>
</tr>
</tbody>
</table>
Fig. 26: Left: The total production cross-section of the strongly interacting particles in comparison with the LHC excluded limits for 7+8 TeV. Here the data from ATLAS and CMS were combined and correspond to the integrated luminosity of 1.3 and 1.1 fb$^{-1}$, respectively. One observes that the cross-section of 0.1 to 0.2 pb is excluded at 95% CL. Right: the cross sections at 14 TeV and expected exclusion for the same limit on the cross-section as at 7 TeV.

Fig. 27: As in Fig. 26, but the excluded region is translated into the $m_{\tilde{g}}, m_{\tilde{q}}$ plane. The red area corresponds to the excluded regions for the integrated luminosity slightly above 1 fb$^{-1}$; the expectations for the higher luminosities have been indicated as well.

component in the gauginos and so the coupling to the weak gauge bosons and Higgs bosons increases, thus increasing the amplitudes for the diagrams in the last row of Fig. 8.

The “strong” production cross sections are characterized by a large number of jets from the long decay chains and the missing energy from the escaping neutralino. These characteristics can be used to efficiently suppress the background. For the electroweak production, both the number of jets and the missing transverse energy is low, since the LSP is not boosted so strongly as in the decay of the heavier strongly interacting particles. Hence, the electroweak gaugino production needs the leptonic decays to reduce the background, so these signatures need more luminosity and cannot compete at present with the sensitivity of the “strong” production of the squarks and gluinos.

The total cross-section for the strongly interacting particles are shown in Fig. 26 together with the excluded region from the direct searches for SUSY particles at the LHC. One observes that the excluded region (below the solid line) follows rather closely the total cross-section, indicated by the colour shading. From the colour coding one observes that the excluded region corresponds to the cross-section limit of about 0.1 – 0.2 pb.
The drop of the excluded region at large values of $m_0$ is due to the fact that in this region the squarks become heavy, which means that the contributions from the diagrams in the second and third rows of Fig. 8 start to diminish. Here only the higher energies will help and doubling the LHC energy from 7 to 14 TeV, as planned in the coming years, quickly increases the cross-section for the gluino production by orders of magnitude, as shown in the right panel of Fig. 26. The expected sensitivity at 14 TeV, plotted as the exclusion contour in case nothing is found, assumes the same efficiency and luminosity (slightly above one fb$^{-1}$ per experiment) as at 7 TeV.

These limits can be translated to the squark and gluino masses as follows. The squark masses have a starting value at the GUT scale equal to $m_0$, but have important contributions from the gluinos in the colour field, so the squark masses are given by $m_{\tilde{q}}^2 \approx m_0^2 + 6.6m_{1/2}^2$, as was determined from the renormalization group equations [30]. Similarly the gluino mass is given by $2.7m_{1/2}$. The term proportional to $m_{1/2}$ in the squark mass corresponds to the self-energy diagrams, which imply that if the "gluino-cloud" is heavy, the squarks cannot be light. This leads to the regions indicated as not allowed ones in Fig. 27. Note that these regions are forbidden in any model with the coupling between the squarks and gluinos, so they are not specific to the CMSSM. The squark masses below 1.1 TeV and the gluino masses below 0.62 TeV are excluded for the LHC data at 7 TeV, as shown in the left panel of Fig. 27. Expected sensitivities for the higher integrated luminosities at 7 and 14 TeV have been indicated as well. One observes that increasing the energy is much more effective than increasing the luminosity. At 14 TeV the squarks with masses of 1.7 TeV and gluinos with masses of 1.02 TeV are within reach with 1 fb$^{-1}$ per experiment, as shown in the right panel of Fig. 27.

9.4 Excluded region for combination of constraints

If one combines the excluded regions from the direct searches at the LHC, the relic density from the WMAP, the already stringent limits on the pseudo-scalar Higgs mass with the XENON100 limits one obtains the excluded region shown in the left panel of Fig. 28. Here the $g-2$ limit is included with the conservative linear addition of theoretical and experimental errors. One observes that the combination excludes $m_{1/2}$ below 525 GeV in the CMSSM for $m_0 < 1500$ GeV, which implies the lower limit on the WIMP mass of 230 GeV and a gluino mass of 1370 GeV, respectively.

As discussed earlier, the LHC becomes rather insensitive to the large $m_0$ region because of the decreasing cross-section for the production of strongly interaction particles and the large background for the production of gauginos. However, in this region one obtains the increased sensitivity above the LHC limits from the relic density and the direct DM searches.
If a Higgs mass of the lightest Higgs boson of 125 GeV is imposed, the preferred region is well above this excluded region, but the size of the preferred region is strongly dependent on the size of the assumed theoretical uncertainty as was shown in Fig. 19. Accepting the 2 GeV uncertainty we get the excluded region shown in Fig. 28 (right panel), which is far above the existing LHC limits and leads to strongly interacting superpartners above 2 TeV. However, in models with an extended Higgs sector, like NMSSM [89], a Higgs mass of 125 GeV can be obtained for lower values of $m_{1/2}$, in which case the regions excluded in the MSSM become viable.

10 The reach of the LHC

10.1 LHC luminosity

The Large Hadron Collider is the unique machine for the new physics searches at the TeV scale. Its c.m. energy is planned to be 14 TeV with very high luminosity up to a few hundred fb$^{-1}$. At the moment the total integrated luminosity in 2012 is already more than 20 fb$^{-1}$. Fig. 29 shows the luminosity delivered in 2012 in $pp$ collisions at the center-of-mass energy of 8 TeV and recorded by ATLAS [90] and CMS [91] experiments.

10.2 Expected LHC reach for SUSY

The LHC is supposed to cover a wide range of parameters of the MSSM (see Figs. below) and discover the superpartners with the masses below 2 TeV. This will be a crucial test for the MSSM and the low energy supersymmetry. The LHC potential to discover supersymmetry is widely discussed in the literature [92, 93].

To present the region of reach for the LHC in different channels of sparticle production it is useful to take the same plane of soft SUSY breaking parameters $m_0$ and $m_{1/2}$. In this case one usually assumes certain luminosity which will be presumably achieved during the accelerator operation. Thus, for instance, in Fig. 30 the regions of reach in different channels are shown. The lines of the constant squark mass form the arch curves, and those for the gluino are almost horizontal. The curved lines show the reach bounds in different channel of creation of the secondary particles. The theoretical curves are obtained within the MSSM for a certain choice of the other soft SUSY breaking parameters. As one can see, for the fortunate circumstances the wide range of the parameter space up to the masses of the order of 2 TeV will be examined. The LHC will be able to discover SUSY with the squark and gluino masses up to $2 \div 2.5$ TeV for the luminosity $L_{\text{tot}} = 100$ fb$^{-1}$. The most powerful signature for the squark and...
Table 4: The background at the hadron colliders: the weak interaction processes (left), and the strong interaction processes (right).

<table>
<thead>
<tr>
<th>Process</th>
<th>Final states</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p\rightarrow\ell\nu\bar{q}$</td>
<td>$2\ell$ $2\bar{j}$ $E_T$</td>
</tr>
<tr>
<td>$p\rightarrow Z\nu\bar{q}$</td>
<td>$\ell$ $2\bar{j}$ $E_T$</td>
</tr>
</tbody>
</table>

Expected sparticle reach in various channels

5 $\sigma$ contours ($\sigma = \frac{N_{sig}}{\sqrt{N_{sig} + N_{bkgd}}}$) for $10^5$pb$^{-1}$

Explorable domain in $\tilde{q}$, $\tilde{g}$ searches in $n$ leptons + $E_T^{miss}$ + $> 2$ jets final states

Fig. 30: The expected range of reach for the superpartners in various channels at the LHC (left) and the expected domain of searches for the squarks and gluinos at the (right) [94].
gluino detection are the multijet events; however, the discovery potential depends on the relation between the LSP, squark, and gluino masses, and decreases with the increase of the LSP mass. The same is true for the sleptons. The typical signal used for the slepton detection is the dilepton pair with the missing energy without hadron jets. For the luminosity of $L_{\text{tot}} = 100 \text{ fb}^{-1}$ the LHC will be able to discover sleptons with the masses up to 400 GeV [92, 93].

### 10.3 Recent results on SUSY searches

Direct searches of the superpartners at the LHC in different channels have pushed the lower limits on their masses, mainly of the gluinos and the squarks of the light two generations, upwards to the TeV range. On the other hand, for the third generation the limits are rather weak and the masses around a few hundred GeV are still allowed. The light third generation squarks are also consistent with the recent observation of the Higgs-like boson with the mass around 125 GeV.

We present here examples on the superparticle searches in various scenarios depicted as exclusion plots. Everywhere in these plots the excluded region is the one below the corresponding curve (lower masses, lower values of parameters).

The first example is the gluino pair production $pp \rightarrow \tilde{g}\tilde{g}$ and $\tilde{g} \rightarrow t\bar{t}\tilde{\chi}_1^0$ decay in the so-called $Gtt$ simplified model. Four different final states (0 leptons with ≥ 3 $b$-jets [95]; 3 leptons with ≥ 4 jets [96]; 0 leptons with ≥ 6-9 jets [97]; and a pair of the same-sign leptons with more than 4 jets [98]) are considered. The first two analysis performed using 12.8 fb$^{-1}$ and 13.0 fb$^{-1}$ data and the last two ones using 5.8 fb$^{-1}$ data. The results slightly differ quantitatively, however, the conclusion is the non-observation of the gluino lighter than 900 GeV (conservative limit) or even 1200 GeV for the lightest neutralino mass less than around 300 GeV.

Another example is the result of searches of the top-squark pair production by ATLAS collaboration based on 4.7 fb$^{-1}$ of $pp$ collision data taken at $\sqrt{S} = 7$ TeV. The exclusion limits at 95% CL are shown in the $t\bar{t} - \tilde{\chi}_1^0$ mass plane. The dashed and solid lines show the expected and observed limits, respectively, including all uncertainties except the theoretical signal cross-section uncertainty (PDF and scale). The dotted lines represent the results obtained when reducing the nominal signal cross-section by...
Fig. 32: Summary of the five dedicated ATLAS searches for the top-squark pair production based on 4.7 fb$^{-1}$ of the $pp$ collision data taken at $\sqrt{s} = 7$ TeV.

Fig. 33: 95% CL exclusion limits for MSUGRA/CMSSM models with $\tan \beta = 10$, $A_0 = 0$ and $\mu > 0$ presented in the $m_0 - m_{1/2}$ plane (left) and in the $m_{\text{gluino}} - m_{\text{squark}}$ plane (right). The blue dashed lines show the expected limits at 95% CL, with the light (yellow) bands indicating the 1$\sigma$ excursions due to experimental uncertainties. The observed limits are indicated by medium (maroon) curves. Previous results from ATLAS [104] are represented by the shaded (light blue) area. The theoretically excluded regions (green and blue) are described in Ref. [105].

1$\sigma$ of its theoretical uncertainty. Depending on the stop mass there can be two different decay channels. For relatively light stops with masses below 200 GeV, the decay $\tilde{t}_1 \to b + \tilde{\chi}_1^{\pm}$, $\tilde{\chi}_1^{\pm} \to W^* \mp \tilde{\chi}_0^0$ is assumed in all the cases, with two hypotheses on the $\tilde{\chi}_1^\pm$, $\tilde{\chi}_1^0$ mass hierarchy, $m(\tilde{\chi}_1^\pm) = 106$ GeV and $m(\tilde{\chi}_1^\pm) = 2m(\tilde{\chi}_1^0)$ [99, 100], see the left panel of Fig. 32. For the heavy stop masses above 200 GeV, the decay $\tilde{t}_1 \to t + \tilde{\chi}_1^0$ is assumed to dominate, the excluded regions are shown in the right panel of Fig. 32 [99–101].

All the exclusion plots discussed above can give direct limits on the masses of supersymmetric particles under certain assumptions (mass relations, dominant decay channels, modified or simplified models, etc.). The latest mass limits for the different models and final state channels obtained by ATLAS are shown in Fig. 34 [106]. Fig. 35 [107, 108] shows the best exclusion limits of the CMS collaboration for 4.98 fb$^{-1}$ data and $\sqrt{s} = 7$ TeV as well as observed limits plotted in the CMSSM $(m_0 - m_{1/2})$ plane.
Is (Low-Energy) SUSY still Alive?

Fig. 34: Mass reach of ATLAS searches for supersymmetry (representative selection)

Fig. 35: Left: Best exclusion limits for the gluino and squark masses, for $m_{\chi^0} = 0$ GeV (dark blue) and $m_{(mother)} - m_{\chi^0} = 200$ GeV (light blue), for each topology, for the hadronic results. Right: Observed limits from several 2011 CMS SUSY searches plotted in the CMSSM ($m_0 - m_{1/2}$) plane.
11 Conclusion

Supersymmetry remains the most popular extension of the Standard Model. Comparison of the MSSM with precision experimental data is as good as for the SM. At the same time, supersymmetry stabilizes the SM due to the cancellation of quadratic divergences to the Higgs boson mass. The prediction of the Higgs boson mass in the MSSM in the region indicated by experimental data can be also considered as an argument in favour of supersymmetry. Besides, the relic density of the DM is not described in the SM but is naturally explained in the MSSM. What is remarkable, the cross section of neutralino annihilation happens to be precisely equal to what is needed for a correct relic density.

Constrained MSSM with a few free parameters seems to satisfy all experimental and theoretical requirements, though recently some tension with the light Higgs boson mass has appeared. The natural way out would be either to release some constraints thus introducing more free parameters or to extend the minimal model, for instance, enlarging the Higgs sector like in the NMSSM. Since it is not clear which model might be correct, all possibilities are open. Unfortunately, there is no "model independent" way of describing SUSY searches, as well as a "smoking gun" process for SUSY except for the discovery of superpartners in the events with missing transverse energy.

Today after 40 years since the invention of supersymmetry we have no single convincing evidence that supersymmetry is realized in particle physics. Still it remains very popular in quantum field theory and in string theory due to its exceptional properties but needs experimental justification.

Let us remind the main pros and contras for supersymmetry in particle physics

Pro:
- Provides natural framework for unification with gravity
- Leads to gauge coupling unification (GUT)
- Solves the hierarchy problem
- Is a solid quantum field theory
- Provides natural candidate for the WIMP cold DM
- Predicts new particles and thus generates new job positions

Contra:
- Does not shed new light on the problem of
  * Quark and lepton mass spectrum
  * Quark and lepton mixing angles
  * the origin of CP violation
  * Number of flavours
  * Baryon asymmetry of the Universe
- Doubles the number of particles

Low energy supersymmetry promises us that new physics is round the corner at the TeV scale to be exploited at colliders and astroparticle experiments of this decade. If our expectations are correct, very soon we will face new discoveries, the whole world of supersymmetric particles will show up and the table of fundamental particles will be enlarged in increasing rate. This would be a great step in understanding the microworld.

Coming back to the question in the title of these lectures, whether SUSY is alive or not, we can say that so far the parameter space of SUSY models is large enough to incorporate all data. Slight tension that appears in particular models can be removed by extension of a model. However, there exist some broad prediction of low energy SUSY that is falsifiable. This is the presence of superpartners at TeV scale. At least some of them should be light enough to be discovered at the LHC at full energy run at 14 TeV. Otherwise, if the scale of SUSY exceeds several TeV, we loose the main arguments in favour of low
energy supersymmetry, namely, the unification of the gauge couplings and the solution of the hierarchy problem. Then the need for a low energy supersymmetry becomes questionable and the possibilities to test it become hardly feasible. The future will show whether we are right in our expectations or not.

Acknowledgements

The authors would like to express their gratitude to the organizers of the School for their effort in creating a pleasant atmosphere and support. This work was partly supported by RFBR grant # 11-02-01177 and Russian Ministry of Education and Science grant # 3802.2012.2.

References


[42] ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaborations, LEP


W. de Boer, M. Huber, A. Gladyshev, D. Kazakov, *The $b \to X(s)\gamma$ decay rate in NLO, Higgs boson limits, and LSP masses in the Constrained Minimal Supersymmetric Model*, hep-ph/0007078, and refs therein


https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResults
https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults
F.E. Paige, SUSY Signatures in ATLAS at LHC, hep-ph/0307342;
D.R. Tovey, Phys. Lett. B498 (2001) 1;
H. Baer, C. Balaz, A. Belyaev, T. Krupovnickas and X. Tata, JHEP 0306 (2003) 054;
S. Sekmen et al., JHEP 1202 (2012) 075;
H. Baer, V. Barger, P. Huang, A. Mustafayev, Phys. Rev. D84 (2011) 091701;
S. Heinemeyer, SUSY Predictions for and from the LHC, arXiv:1103.0952 [hep-ph];
M. Dittmar, SUSY discovery strategies at the LHC, hep-ex/9901004;
http://CMSinfo.cern.ch/Welcome.html-/CMSdocuments/CMSplots
ATLAS collaboration, Search for gluino pair production in events with missing transverse momentum and at least three b-jets using 13.0 fb$^{-1}$ of pp Collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector, ATLAS-CONF-2012-145.
ATLAS collaboration, Search for supersymmetry in events with three leptons, multiple jets, and missing transverse momentum in 13.0 fb$^{-1}$ of pp collisions with the ATLAS detector at $\sqrt{s} = 8$ TeV, ATLAS-CONF-2012-151.
Is (Low-Energy) SUSY still Alive?

[97] ATLAS collaboration, *Search for new phenomena using large jet multiplicities and missing transverse momentum with ATLAS in 5.8 fb$^{-1}$ of $\sqrt{s}=8$ TeV proton-proton collisions*, ATLAS-CONF-2012-103.


[100] ATLAS collaboration, *Search for light top squark pair production in final states with leptons and b-jets with the ATLAS detector in $\sqrt{s}=7$ TeV proton-proton collisions*, arxiv:1209.2102


[106] https://twiki.cern.ch/twiki/bin/view/AtlasPublic/CombinedSummaryplots


[108] https://twiki.cern.ch/twiki/bin/view/CMSPublic/SUSYSMSummaryplots
Neutrino Physics

G. Barenboim
Universitat de València – CSIC, València, Spain

Abstract
The Standard Model has been incredibly successful in predicting the outcome of almost all the experiments done up so far. In it, neutrinos are mass-less. However, in recent years we have accumulated evidence pointing to tiny masses for the neutrinos (as compared to the charged leptons). These masses allow neutrinos to change their flavour and oscillate. In these lectures I review the properties of neutrinos in and beyond the Standard Model.

1 Introduction
The last decade witnessed a revolution in neutrino physics. It has been observed that neutrinos have nonzero masses, and that leptons mix. This fact was proven by the observation that neutrinos can change from one type, or “flavour”, to another. Almost all the knowledge we have gathered about neutrinos, is only fifteen years old. But before diving into the recent "news" about neutrinos, lets find out how neutrinos were born.

The 20s witnessed the assassination of many sacred cows, and physics was no exception, one of physic’s most holy principles, the conservation of energy, appeared not to hold within the subatomic world. For some radioactive nuclei, it seemed that part of its energy just disappear, leaving no footprint of its existence.

In 1920, in a letter to a conference, Pauli wrote, "Dear radioactive Ladies and Gentlemen, ... as a desperate remedy to save the principle of energy conservation in beta decay, ... I propose the idea of a neutral particle of spin half”. Pauli postulated that the energy loss was taken off by a new particle, whose properties were such that it would not yet be seen: it had no electric charge and rarely interacted with matter at all. This way, the neutrino was born into the world of particle physics.

Soon afterwards, Fermi wrote the four-Fermi Hamiltonian for beta decay using the neutrino, electron, neutron and proton. A new field came to existence: the field of weak interactions. And two decades after Pauli’s letter, Cowan and Reines finally observed anti-neutrinos emitted by a nuclear reactor. As more and more particles were discovered in the following years and observed to participate in weak processes, weak interactions got legitimacy as a new force of nature and the neutrino became a key ingredient of this interactions.

Further experiments over the course of the next 30 years showed us that there were three kinds, or “flavours” of neutrinos (electron neutrinos ($\nu_e$), muon neutrinos ($\nu_\mu$) and tau neutrinos ($\nu_\tau$)) and that, as far as we could tell, had no mass at all. The neutrino saga might have stop there, but new experiments in solar physics taught us that the neutrino story was just beginning....

Within the Standard Model, neutrinos have zero mass and therefore interact diagonally in flavour space,

\[
W^+ \rightarrow e^+ + \nu_e \quad ; \quad Z \rightarrow \nu_e + \bar{\nu}_e \\
W^+ \rightarrow \mu^+ + \nu_\mu \quad ; \quad Z \rightarrow \nu_\mu + \bar{\nu}_\mu \\
W^+ \rightarrow \tau^+ + \nu_\tau \quad ; \quad Z \rightarrow \nu_\tau + \bar{\nu}_\tau
\]  

(1)

Since they are mass-less, they move at the speed of light and therefore their flavour remains the same from production up to detection. It is obvious then, that at least as flavour is concerned, zero mas neutrinos are almost not interesting as compared to quarks.
On the other hand, neutrinos masses different from zero, mean that there are three neutrino mass eigenstates $\nu_i, i = 1, 2, \ldots$, each with a mass $m_i$. The meaning of leptonic mixing can be understood by analysing the leptonic decays, $W^+ \rightarrow \nu_i + \ell_\alpha$ of the charged $W$ boson. Where, $\alpha = e, \mu, \tau$, and $\ell_e$ represents the electron, $\ell_\mu$ the muon, or $\ell_\tau$ the tau. We refer to particle $\ell_\alpha$ as the charged lepton of flavour $\alpha$. Mixing essentially means that when the $W^+$ decays to a given flavour of charged lepton $\ell_\alpha$, the neutrino that comes along is not always the same mass eigenstate $\nu_i$. Any of the different $\nu_i$ can show up. The amplitude for the decay of a $W^+$ to a specific combination $\overline{\nu}_\alpha + \nu_i$ is designated by $U^*_{\alpha i}$. The neutrino that is emitted in $W^+$ decay along with the given charged lepton $\ell_\alpha$ is then

$$|\nu_\alpha >= \sum_i U^*_{\alpha i} |\nu_i > . \quad (2)$$

This particular combination of mass eigenstates is the neutrino of flavour $\alpha$.

The quantities $U_{\alpha i}$ can be collected in a unitary matrix (analogue to the CKM matrix of the quark sector) known as the leptonic mixing matrix \cite{1}. The unitarity of $U$ guarantees that every time a neutrino of flavour $\alpha$ interacts in a detector and produces a charged lepton, such a charged lepton will be always $\ell_\alpha$, the charged lepton with flavour $\alpha$. That is, a $\nu_e$ indelibly creates an $e$, a $\nu_\mu$ a $\mu$, and a $\nu_\tau$ a $\tau$.

The relation (2), describing a neutrino of definite flavour as a linear combination of mass eigenstates, may be inverted to describe each mass eigenstate $\nu_i$ as a linear combination of flavours:

$$|\nu_i > = \sum_\alpha U_{\alpha i} |\nu_\alpha > . \quad (3)$$

The $\alpha$-flavour "content" (or fraction) of $\nu_i$ is clearly $|U_{\alpha i}|^2$. When a $\nu_i$ interacts and generates a charged lepton, this $\alpha$-flavour fraction becomes the probability that the emerging charged lepton be of flavour $\alpha$.

2 Neutrino oscillations in vacuum

A standard neutrino flavour transition, or "oscillation", can be understood as follows. A neutrino is produced by a source together with a charged lepton $\overline{\nu}_\alpha$ of flavour $\alpha$. Therefore, at the production point, the neutrino is a $\nu_\alpha$. Then, after birth, the neutrino travels a distance $L$ until it is detected. There, it is where it reaches a target with which it interacts and produces another charged lepton $\ell_\beta$ of flavour $\beta$. Thus, at the interaction point, the neutrino is a $\nu_\beta$. If $\beta \neq \alpha$ (for example, if $\ell_\alpha$ is a $\mu$ but $\ell_\beta$ is a $\tau$), then, during its trip from the source to the detection point, the neutrino has transitioned from a $\nu_\alpha$ into a $\nu_\beta$.

This morphing of neutrino flavour, $\nu_\alpha \rightarrow \nu_\beta$, is a text-book example of a quantum-mechanical effect.

Because, as described by Eq. (2), a $\nu_\alpha$ is really a coherent superposition of mass eigenstates $\nu_i$, the neutrino that propagates since it is created until it interacts, can be any one of the $\nu_i$’s, therefore we must add the contributions of all the different $\nu_i$ coherently. Then, the transition amplitude, $\text{Amp}(\nu_\alpha \rightarrow \nu_\beta)$ contains a share of each $\nu_i$ and it is a product of three factors. The first one is the amplitude for the neutrino born at the production point in combination with an $\overline{\nu}_\alpha$ to be, specifically, a $\nu_\alpha$. As we have mentioned already , this amplitude is given by $U^*_{\alpha i}$. The second factor is the amplitude for the $\nu_i$ created by the source to propagate until it reaches the detector. We will call this factor $\text{Prop}(\nu_i)$ and will find out its value later. The third factor is the amplitude for the charged lepton produced by the interaction of the $\nu_i$ with the detector to be, specifically, an $\ell_\beta$. As the Hamiltonian that describes the interaction of neutrinos, charged leptons and $W$ bosons is hermitian , it ensues that if $\text{Amp}(W \rightarrow \overline{\nu}_\alpha \nu_i) = U^*_{\alpha i}$, then $\text{Amp}(\nu_i \rightarrow \ell_\beta W) = U_{\beta i}$. Therefore, the third and last factor in the $\nu_i$ contribution is $U_{\beta i}$, and

$$\text{Amp}(\nu_\alpha \rightarrow \nu_\beta) = \sum_i U^*_{\alpha i} \text{Prop}(\nu_i) U_{\beta i} . \quad (4)$$
It still remains to be established the value of $\text{Prop}(\nu_i)$. To determine it, we’d better study the $\nu_i$ in its rest frame. We will label the time in that system $\tau_i$. If $\nu_i$ does have a rest mass $m_i$, then in this frame its state vector satisfies the good old Schrödinger equation
\[ i \frac{\partial}{\partial \tau_i} |\nu_i(\tau_i)\rangle = m_i |\nu_i(\tau_i)\rangle . \] (5)
whose solution is given clearly by
\[ |\nu_i(\tau_i)\rangle = e^{-i m_i \tau_i} |\nu_i(0)\rangle . \] (6)

Then, the amplitude for the mass eigenstate $\nu_i$ to propagate for a time $\tau_i$, is simply the amplitude $< \nu_i(0)|\nu_i(\tau_i)\rangle$ for observing the original $\nu_i |\nu_i(0)\rangle$ after some time as the evolved state $|\nu_i(\tau_i)\rangle$, i.e. $\exp[-i m_i \tau_i]$. Thus $\text{Prop}(\nu_i)$ is only this amplitude where we have used that the time taken by $\nu_i$ to travel from the neutrino source to the detector is $\tau_i$, the proper time.

Nevertheless, if we want $\text{Prop}(\nu_i)$ to be of any use to us, we must write it first in terms of variables in the laboratory system. The natural choice is obviously the laboratory-frame distance, $L$, that the neutrino covers between the source and the detector, and the laboratory-frame time, $t$, that slips away during the journey. The distance $L$ is set by the experimentalists through the selection of the place of settlement of the source and that of the detector. Likewise, the value of the time $t$ is selected by the experimentalists through their election for the time at which the neutrino is created and that when it is detected. Thus, $L$ and $t$ are chosen (hopefully carefully) by the experiment design, and are the same for all the $\nu_i$ components of the beam. Different $\nu_i$ do travel through an identical distance $L$, in an identical time $t$.

We still need two other laboratory-frame variables, they are the laboratory-frame energy $E_i$ and momentum $p_i$ of the neutrino mass eigenstate $\nu_i$. With the four lab-frame variable and using Lorentz invariance, we can obtain the phase $m_i \tau_i$ in the $\nu_i$ propagator $\text{Prop}(\nu_i)$ we have been looking for, which (expressed in terms of laboratory frame variables) is given by
\[ m_i \tau_i = E_i t - p_i L . \] (7)

At this point however one may argue that, in real life, neutrino sources are basically constant in time, and that the time $t$ that slips away since the neutrino is produced till it dies in the detector is actually not measured. This argument is right. In real life, an experiment averages over the time $t$ used by the neutrino to complete its journey. However, lets consider that two constituents of the neutrino beam, the first one with energy $E_1$ and the second one with energy $E_2$ (both measured in the lab frame), contribute
coherently to the neutrino signal produced in the detector. Now, if we call $t$ to the time used by the neutrino to cover the distance separating the production and detection points, then by the time the constituent whose energy is $E_j$ ($j = 1, 2$) arrives to the detector, it has raised a phase factor $\exp[-iE_j t]$. Therefore, we will have an interference between the $E_1$ and $E_2$ beam participants that will include a phase factor $\exp[-i(E_1 - E_2)t]$. When averaged over the non-observed travel time $t$, this factor goes away, except when $E_2 = E_1$. Therefore, only components of the neutrino beam that share the same energy contribute coherently to the neutrino oscillation signal [2, 3]. Specifically, only the different mass eigenstate constituents of a beam that have the same energy contribute coherently to the oscillation signal.

A mass eigenstate $\nu_i$, with mass $m_i$, and energy $E$, has a momentum $p_i$ given by

$$p_i = \sqrt{E^2 - m_i^2} \cong E - \frac{m_i^2}{2E}. \quad (8)$$

Where, we have used that as the masses of the neutrinos are miserably small, $m_i^2 \ll E^2$ for any energy $E$ attainable at a realistic experiment. From Eqs. (7) and (8), we see that at energy $E$ the phase $m_i \tau_i$ appearing in $\text{Prop}(\nu_i)$ takes the value

$$m_i \tau_i \cong E(t - L) + \frac{m_i^2 L}{2E}. \quad (9)$$

As the phase $E(t - L)$ appears in all the interfering terms it will eventually disappear when calculating the transition amplitude. Thus, we can get rid of it already now and use

$$\text{Prop}(\nu_i) = \exp[-im_i^2 \frac{L}{2E}]. \quad (10)$$

Applying this result, we can obtain from Eq. (4) that the amplitude for a neutrino born as a $\nu_\alpha$ to be detected as a $\nu_\beta$ after covering a distance $L$ through vacuum with energy $E$ yields

$$\text{Amp}(\nu_\alpha \rightarrow \nu_\beta) = \sum_i U_{\alpha i}^* e^{-im_i^2 \frac{L}{2E}} U_{\beta i}. \quad (11)$$

The expression above is valid for an arbitrary number of neutrino flavours and mass eigenstates. The probability $P(\nu_\alpha \rightarrow \nu_\beta)$ for $\nu_\alpha \rightarrow \nu_\beta$ can be found by squaring it, giving

$$P(\nu_\alpha \rightarrow \nu_\beta) = |\text{Amp}(\nu_\alpha \rightarrow \nu_\beta)|^2$$

$$= \delta_{\alpha\beta} - 4 \sum_{i > j} \Re(U_{\alpha i}^* U_{\beta i} U_{\alpha j} U_{\beta j}^*) \sin^2 \left( \frac{\Delta m_{ij}^2 L}{4E} \right)$$

$$+ 2 \sum_{i > j} \Im(U_{\alpha i}^* U_{\beta i} U_{\alpha j} U_{\beta j}^*) \sin \left( \frac{\Delta m_{ij}^2 L}{2E} \right), \quad (12)$$

with

$$\Delta m_{ij}^2 \equiv m_i^2 - m_j^2. \quad (13)$$

In order to get Eq. (12) we have used that the mixing matrix $U$ is unitary.

The oscillation probability $P(\nu_\alpha \rightarrow \nu_\beta)$ we have just obtained corresponds to that of a neutrino, and not to an antineutrino, as we have used that the oscillating neutrino was produced along with a charged antilepton $\bar{\ell}$, and gives birth to a charged lepton $\ell$ once it reaches the detector. The corresponding probability $P(\bar{\nu}_\alpha \rightarrow \bar{\nu}_\beta)$ for an antineutrino oscillation can be obtained from $P(\nu_\alpha \rightarrow \nu_\beta)$ taking advantage of the fact that the two transitions $\bar{\nu}_\alpha \rightarrow \bar{\nu}_\beta$ and $\nu_\beta \rightarrow \nu_\alpha$ are CPT conjugated processes. Thus, assuming that neutrino interactions respect CPT [4],

$$P(\bar{\nu}_\alpha \rightarrow \bar{\nu}_\beta) = P(\nu_\beta \rightarrow \nu_\alpha). \quad (14)$$
Then, from Eq. (12) we obtain that
\[ P(\nu_\beta \rightarrow \nu_\alpha; U) = P(\nu_\alpha \rightarrow \nu_\beta; U^*) . \]
(15)
Therefore, if CPT is a good symmetry (with respect to neutrino interactions), Eq. (12) tells us that
\[ P(\nu_\beta \rightarrow \nu_\alpha; U) = \delta_{\alpha\beta} - 4 \sum_{i>j} \Re(U^*_{\alpha i}U_{\beta i}U_{\alpha j}U^*_{\beta j}) \sin^2 \left( \frac{\Delta m^2_{ij}L}{4E} \right) \]
\[ + \frac{1}{\epsilon} \sum_{i>j} \Im(U^*_{\alpha i}U_{\beta i}U_{\alpha j}U^*_{\beta j}) \sin \left( \frac{\Delta m^2_{ij}L}{2E} \right) . \]
(16)
These expressions make it clear that if the mixing matrix \( U \) is complex, \( P(\nu_\alpha \rightarrow \nu_\beta) \) and \( P(\nu_\alpha \rightarrow \nu_\beta) \) will not be identical, in general. As \( \nu_\alpha \rightarrow \nu_\beta \) and \( \nu_\alpha \rightarrow \nu_\beta \) are CP conjugated processes, \( P(\nu_\alpha \rightarrow \nu_\beta) \neq P(\nu_\alpha \rightarrow \nu_\beta) \) would be an evidence of CP violation in neutrino oscillations. So far, CP violation has been observed only in the quark sector, so its measurement in neutrino oscillations would be quite exciting.

So far, we have been working in natural units, if we return now the \( h \)'s and \( e \) factor (we have happily left out) into the oscillation probability we find that
\[ \sin^2 \left( \frac{\Delta m^2_{ij}L}{4E} \right) \rightarrow \sin^2 \left( \frac{\Delta m^2_{ij}c^4 L}{4\hbar c E} \right) \]
(17)
Having done that, it is easy and instructive to explore the semi-classical limit, \( h \rightarrow 0 \). In this limit the oscillation length goes to zero (the oscillation phase goes to infinity) and the oscillations are averaged out. The same happens if we let the mass difference \( \Delta m^2 \) become large. This is exactly what happens in the quark sector (and the reason why we never study quark oscillations despite knowing that mass eigenstates do not coincide with flavour eigenstates).

In terms of real life units (which are not "natural" units), the oscillation phase is given by
\[ \Delta m_{ij}^2 \frac{L}{4E} = 1.27 \Delta m_{ij}^2(eV^2) \frac{L}{E(GeV)} \]
(18)
then, since \( \sin^2[1.27 \Delta m_{ij}^2(eV^2)L/(km)/E(GeV)] \) can be experimentally observed only if its argument is of order unity or larger, an experimental set-up with a baseline \( L \) (km) and an energy \( E \) (GeV) is sensitive to neutrino mass squared differences \( \Delta m_{ij}^2(eV^2) \) larger that or equal to \( \sim [L(km)/E(GeV)]^{-1} \). For example, an experiment with \( L \sim 10^4 \) km, roughly the size of Earth’s diameter, and \( E \sim 1 \) GeV is sensitive to \( \Delta m_{ij}^2 \) down to \( \sim 10^{-4} \) eV\(^2\). This fact makes it clear that neutrino oscillation experiments can test super tiny neutrino masses. It does so by exhibiting quantum mechanical interferences between amplitudes whose relative phases are proportional to these super tiny neutrino mass squared differences, which can be transformed into sizeable effects by choosing an \( L/E \) large enough.

But let’s keep analysing the oscillation probability and see whether we can learn more about neutrino oscillations by studying its expression.

It is clear from \( P(\nu_\alpha \rightarrow \nu_\beta) \) that if neutrinos have zero mass, in such a way that all \( \Delta m_{ij}^2 = 0 \), then, \( P(\nu_\alpha \rightarrow \nu_\beta) = \delta_{\alpha\beta} \). Therefore, the experimental observation that neutrinos can morph from one flavour to a different one indicates that neutrinos are not only massive but also that their masses are not degenerate. Actually, it was precisely this evidence the one that led to the conclusion that neutrinos are massive.

However, every neutrino oscillation seen so far has involved at some point neutrinos that travel through matter. But the expression we derived is valid only for flavour change in vacuum, and does not take into account any interaction between the neutrinos and the matter traversed between their source.
and their detector. Thus, one might ask whether flavour-changing interactions between neutrinos and matter are indeed responsible of the observed flavour changes, and not neutrino masses. Regarding this question, two points can be made. First, although it is true that the Standard Model of elementary particle physics contains only mass-less neutrinos, it provides an amazingly well corroborated description of neutrino interactions, and this description clearly establishes that neutrino interactions with matter do not change flavour. Second, for at least some of the observed flavour changes, matter effects are expected to be miserably small, and there is solid evidence that in these cases, the flavour transition probability depends on $L$ and $E$ through the combination $L/E$, as anticipated by the oscillation hypothesis. Modulo a constant, $L/E$ is precisely the proper time that goes by in the rest frame of the neutrino as it covers a distance $L$ possessing an energy $E$. Thus, these flavour transitions behave as if they were a progression of the neutrino itself over time, rather than a result of interaction with matter.

Now, suppose the leptonic mixings were trivial. This would mean that in the decay $W^+ \rightarrow \ell_\alpha + \nu_\alpha$, which as we established has an amplitude $U_{\alpha i}^*$, the emerging charged antilepton $\ell_\alpha$ of flavour $\alpha$ comes along always with the same neutrino mass eigenstate $\nu_i$. That is, if $U_{\alpha i}^* \neq 0$, then $U_{\alpha j}$ becomes zero for all $j \neq i$. Therefore, from Eq. (16) it is clear that, $P(\bar{\nu}_\alpha \rightarrow \bar{\nu}_\beta) = \delta_{\alpha\beta}$. Thus, the observation that neutrinos can change flavour indicates mixing.

Then, there are basically two ways to detect neutrino flavour change. The first one is to observe, in a beam of neutrinos which are all created of the same flavour, say $\alpha$, some appearance of neutrinos of a new flavour $\beta$ that is different from the flavour $\alpha$ we started with. This is what is called an appearance experiment. The second way is to start with a beam of identical $\nu_\alpha$s, whose flux is known, and observe that this known $\nu_\alpha$ flux is depleted. This is called a disappearance experiment.

As Eq. (16) shows, the transition probability in vacuum does not only depend on $L/E$ but also oscillates with it. It is because of this fact that neutrino flavour transitions are named “neutrino oscillations”. Now notice also that neutrino transition probabilities depend only on neutrino squared-mass splittings, and not on the individual squared neutrino masses themselves. Thus, oscillation experiments can only measure the neutrino squared-mass spectral pattern, but not its absolute scale, i.e. the distance above zero the entire pattern lies.

It is clear that neutrino transitions cannot modify the total flux in a neutrino beam, but simply alter its distribution between the different flavours. Actually, from Eq. (16) and the unitarity of the $U$ matrix, it follows that

$$
\sum_{\beta} P(\bar{\nu}_\alpha \rightarrow \bar{\nu}_\beta) = 1 ,
$$

where the sum runs over all flavours $\beta$, including the original flavour $\alpha$. Eq. (19) makes it transparent that the probability that a neutrino morphs its flavour, added to the probability that it does not do so, is one. Ergo, flavour transitions do not change the total flux. Nevertheless, some of the flavours $\beta \neq \alpha$ into which a neutrino can oscillate into may be sterile flavours; that is, flavours that do not take part in weak interactions and therefore escape detection. If any of the original (active) neutrino flux turns into sterile, then an experiment measuring the total active neutrino flux—that is, the sum of the $\nu_e$, $\nu_\mu$, and $\nu_\tau$ fluxes—will find it to be less than the original flux. In the experiments performed up today, no flux was ever missed.

In the literature, description of neutrino oscillation normally assume that the different mass eigenstates $\nu_i$ that contribute coherently to a beam share the same momentum, rather than the same energy as we have argued they must have. While the supposition of equal momentum is technically wrong, it is an inoffensive mistake, since, as can easily be shown [5], it conveys to the same oscillation probabilities as we have found.

A relevant and interesting case of the (not that simple) formula for $P(\bar{\nu}_\alpha \rightarrow \bar{\nu}_\beta)$ is the case where only two flavours participate in the oscillation. The only-two-neutrino scenario is a rather rigorous description of a vast number of experiments. Let's assume then, that only two mass eigenstates, which we will name $\nu_1$ and $\nu_2$, and two corresponding flavour states, which we will name $\nu_\mu$ and $\nu_\tau$, are relevant.
There is then only one squared-mass splitting, $m_2^2 - m_1^2 \equiv \Delta m^2$. Even more, neglecting phase factors that can be proven to have no effect on oscillation, the mixing matrix $U$ takes the simple form

$$
\begin{pmatrix}
    \nu_\mu \\
    \nu_\tau
\end{pmatrix}
= \begin{pmatrix}
    \cos \theta & \sin \theta \\
    -\sin \theta & \cos \theta
\end{pmatrix}
\begin{pmatrix}
    \nu_1 \\
    \nu_2
\end{pmatrix}
$$

The $U$ of Eq. (20) is just a $2 \times 2$ rotation matrix, and the rotation angle $\theta$ within it is referred to as the mixing angle. Inserting the $U$ of Eq. (20) and the unique $\Delta m^2$ into the general expression for $P(\nu_\alpha \rightarrow \nu_\beta)$, Eq. (16), we immediately find that, for $\beta \neq \alpha$, when only two neutrinos are relevant,

$$
P(\nu_\alpha \rightarrow \nu_\beta) = \sin^2 2\theta \sin^2 \left( \frac{\Delta m^2 L}{4E} \right).
$$

Moreover, the survival probability, i.e., the probability that the neutrino remains with the same flavour it was created with, is, as usual, unity minus the probability that it changes flavour.

### 3 Neutrino oscillations in matter

When we create a beam of neutrinos on earth through an accelerator and send it up to thousand kilometres away to a meet detector, the beam does not move through vacuum, but through matter, earth matter. The beam of neutrinos then scatters (coherently forward) from particles it meets along the way. Such a scattering can have a large effect on the transition probabilities. We will assume that neutrino interactions with matter are flavour conserving, as described by the Standard Model. Then a neutrino in matter have two possibilities to enjoy coherent forward scattering from matter particles. First, if it is an electron neutrino, $\nu_e$—and only in this case—can exchange a $W$ boson with an electron. Neutrino-electron coherent forward scattering via $W$ exchange opens up an extra interaction potential energy $V_W$ suffered exclusively by electron neutrinos. Obviously, this additional weak interaction energy has to be proportional to $G_F$, the Fermi coupling constant. In addition, the interaction energy coming from $\nu_e - e$ scattering grows with $N_e$, the number of electrons per unit volume. From the Standard Model, we find that

$$
V_W = +\sqrt{2}G_FN_e,
$$

clearly, this interaction energy affects also antineutrinos (in a opposite way though), it changes sign if we replace the $\nu_e$ by $\bar{\nu}_e$.

The second interaction corresponds to the case where a neutrino in matter exchanges a $Z$ boson with an matter electron, proton, or neutron. The Standard Model teaches us that weak interactions are flavour blind. Every flavour of neutrino enjoys them, and the amplitude for this $Z$ exchange is always the same. It also teaches us that, at zero momentum transfer, the $Z$ couplings to electrons and protons have equal strength and opposite sign. Therefore, counting on the fact that the matter through which our neutrino moves is electrically neutral (it contains equal number of electrons and protons), the electron and proton contribution to coherent forward neutrino scattering through $Z$ exchange will add up to zero. Then, the effect of the $Z$ exchange contribution to the interaction potential energy $V_Z$ will be equal to all flavors and will depends exclusively on $N_n$, the number density of neutrons. From the Standard Model, we find that

$$
V_Z = -\frac{\sqrt{2}}{2}G_FN_n,
$$

as was the case before, for $V_W$, this contribution will flip sign if we replace the neutrinos by antineutrinos.

But we already learnt that Standard Model interactions do not change neutrino flavour. Therefore, unless non-Standard-Model flavour changing interactions play a role, neutrino flavour transitions or neutrino oscillations points also to neutrino mass and mixing even when neutrinos are propagating through matter.
Neutrino propagation in matter is easy to understand when analyzed through a time dependent Schrödinger equation in the laboratory frame

\[ i \hbar \frac{\partial}{\partial t} |\nu(t)\rangle = \mathcal{H} |\nu(t)\rangle \]  

(24)

where, $|\nu(t)\rangle$ is a multicomponent neutrino vector state, in which each neutrino flavour corresponds to one component. In the same way, the Hamiltonian $\mathcal{H}$ is a matrix in flavour space. To make our lives easy, let’s analyze the case where only two neutrino flavours are relevant, say $\nu_e$ and $\nu_\mu$. Then

\[ |\nu(t)\rangle = \begin{pmatrix} f_e(t) \\ f_\mu(t) \end{pmatrix} , \]

(25)

where $f_e(t)^2$ is the fraction of the neutrino that is a $\nu_e$ at time $t$, and similarly for $f_\mu(t)$. Analogously, $\mathcal{H}$ is a 2×2 matrix in $\nu_e - \nu_\mu$ space.

It will prove to be clarifying to work out the two-flavour case in vacuum first, and add matter effects afterwards. Using Eq. (2) for $|\nu_\alpha\rangle$ as a linear combination of mass eigenstates, we can see that the $\nu_\alpha - \nu_\beta$ matrix element of the Hamiltonian in vacuum, $\mathcal{H}_{\text{Vac}}$, can be written as

\[ <\nu_\alpha|\mathcal{H}_{\text{Vac}}|\nu_\beta> = \sum_j U_{\alpha j}^* U_{\beta j} |\nu_j\rangle \]  

\[ = \sum_j U_{\alpha j} U_{\beta j}^* \sqrt{p^2 + m_j^2} . \]

(26)

where we are supposing that neutrino belongs to a beam where all its mass components (the mass eigenstates) share the same definite momentum $p$. (As we have already mentioned, this supposition is technically wrong, however it leads anyway to the right oscillation probability.) In the second line of Eq. (26), we have used that

\[ \mathcal{H}_{\text{Vac}}|\nu_j\rangle = E_j|\nu_j\rangle \]  

(27)

with $E_j = \sqrt{p^2 + m_j^2}$ the energy of the mass eigenstate $\nu_j$ with momentum $p$, and the fact that the mass eigenstates of the Hermitian Hamiltonian $\mathcal{H}_{\text{Vac}}$ constitute a basis and therefore are orthogonal.

As we have already mentioned, neutrino oscillations are the archetype quantum interference phenomenon, where only the relative phases of the interfering states play a role. Therefore, only the relative energies of these states, which set their relative phases, are relevant. As a consequence, if it proves to be convenient, we can feel free to happily remove from the Hamiltonian $\mathcal{H}$ any contribution proportional to the identity matrix $I$. As we have said, this subtractions will leave unaffected the differences between the eigenvalues of $\mathcal{H}$, and therefore will leave unaffected the prediction of $\mathcal{H}$ for flavour transitions.

It goes without saying that as in this case only two neutrinos are relevant, there are only two mass eigenstates, $\nu_1$ and $\nu_2$, and only one mass splitting $\Delta m^2 \equiv m_2^2 - m_1^2$, and therefore the $U$ matrix is given by Eq. (20). Inserting this matrix into Eq. (26), and applying the high momentum approximation $(p^2 + m_j^2)^{1/2} \approx p + m_j^2/2p$, and removing from $\mathcal{H}_{\text{Vac}}$ a term proportional to the identity matrix (a removal we know is going to be harmless), we get

\[ \mathcal{H}_{\text{Vac}} = \frac{\Delta m^2}{4E} \begin{pmatrix} -\cos 2\theta & \sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{pmatrix} . \]  

(28)

To write this expression, we have used that $p \approx E$, where $E$ is the average energy of the neutrino mass eigenstates in our neutrino beam of ultra high momentum $p$.

It is not difficult to corroborate that the Hamiltonian $\mathcal{H}_{\text{Vac}}$ of Eq. (28) for the two neutrino scenario would give an identical oscillation probability, Eq. (21), as the one we have already obtained in a
different way. For example, let's have a look at the oscillation probability for the process $\nu_e \rightarrow \nu_\mu$. From Eq. (20) it is clear that in terms of the mixing angle, the electron neutrino state composition is

$$ |\nu_e \rangle = |\nu_1 \rangle \cos \theta + |\nu_2 \rangle \sin \theta , $$

(29)

while that of the muon neutrino is given by

$$ |\nu_\mu \rangle = -|\nu_1 \rangle \sin \theta + |\nu_2 \rangle \cos \theta . $$

(30)

Now, the eigenvalues of $H_{\text{Vac}}$, Eq.25, read

$$ \lambda_1 = -\frac{\Delta m^2}{4E}, \quad \lambda_2 = +\frac{\Delta m^2}{4E}. $$

(31)

The eigenvectors of this Hamiltonian, $|\nu_1 \rangle$ and $|\nu_2 \rangle$, can also be written in terms of $|\nu_e \rangle$ and $|\nu_\mu \rangle$ by means of Eqs. (29) and (30). Therefore, with $H$, its vacuum expression, $H_{\text{Vac}}$ of Eq. (28), the Schrödinger equation of Eq. (24) tells us that if at time $t = 0$ we begin from a $|\nu_e \rangle$, then after some time $t$ this $|\nu_e \rangle$ will progress into the state

$$ |\nu(t) \rangle = |\nu_1 \rangle e^{\frac{i\Delta m^2}{4E} t} \cos \theta + |\nu_2 \rangle e^{-\frac{i\Delta m^2}{4E} t} \sin \theta . $$

(32)

Thus, the probability $P(\nu_e \rightarrow \nu_\mu)$ that this evolved neutrino be detected as a different flavor $\nu_\mu$, from Eqs. (30) and (32), is given by,

$$ P(\nu_e \rightarrow \nu_\mu) = |<\nu_\mu|\nu(t)\rangle|^2 = |\sin \theta \cos \theta (-e^{\frac{i\Delta m^2}{4E} t} + e^{-\frac{i\Delta m^2}{4E} t})|^2 = \sin^2 \theta \sin^2 \left( \frac{\Delta m^2 L}{4E} \right). $$

(33)

Where we have substituted the time $t$ travelled by our highly relativistic state by the distance $L$ it has covered. The flavour transition or oscillation probability of Eq. (33), as expected, is exactly the same we have found before, Eq. (21).

We can now move on to analyze neutrino propagation in matter. In this case, the $2 \times 2$ vacuum Hamiltonian $H_{\text{Vac}}$ receives two additional contributions and becomes $H_M$, which is given by

$$ H_M = H_{\text{Vac}} + V_W \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + V_Z \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. $$

(34)

In the new Hamiltonian, the first additional contribution corresponds to the interaction potential due to $W$ exchange, Eq. (22). As this interaction is suffered only by $\nu_e$, this contribution is different from zero only in the upper left, $\nu_e - \nu_e$, element of $H_M$. The second additional contribution, the last term of Eq. (34) comes from the interaction potential due to $Z$ exchange, Eq. (23). Since this interaction is flavour blind, it affects every neutrino flavour in the same way, its contribution to $H_M$ is proportional to the identity matrix, and can be safely neglected. Then

$$ H_M = H_{\text{Vac}} + \frac{V_W}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, $$

(35)

where (for reasons that are going to become clear later) we have divided the $W$-exchange contribution into two pieces, one a multiple of the identity matrix (that we will disregard in the next step) and, a piece that it is not a multiple of the identity matrix. Disregarding the first piece as promised, we have from Eqs. (28) and (35)

$$ H_M = \frac{\Delta m^2}{4E} \begin{pmatrix} -(\cos 2\theta - A) & \sin 2\theta \\ \sin 2\theta & (\cos 2\theta - A) \end{pmatrix}, $$

(36)
where
\[ A \equiv \frac{V_W/2}{\Delta m^2/4E} = \frac{2\sqrt{2} G_F N_e E}{\Delta m^2} \]  

Clearly, \( A \) shows the size (the importance) of the matter effects as compared to the neutrino squared-mass splitting and signal the situations when they become important.

Now, if we define
\[ \Delta m^2_M \equiv \Delta m^2 \sqrt{\sin^2 2\theta + (\cos 2\theta - A)^2} \]

and
\[ \sin^2 2\theta^M = \frac{\sin^2 2\theta}{\sin^2 2\theta + (\cos 2\theta - A)^2} \]

\( \mathcal{H}_M \) can be given as
\[ \mathcal{H}_M = \frac{\Delta m^2_M}{4E} \left( -\cos 2\theta^M \sin 2\theta^M \cos 2\theta^M \right) \]

That is, the Hamiltonian in matter, \( \mathcal{H}_M \), becomes formally identical to its vacuum counterpart, \( \mathcal{H}_{\text{Vac}} \), Eq. (28), except that the vacuum parameters \( \Delta m^2 \) and \( \theta \) are now given by the matter ones, \( \Delta m^2_M \) and \( \theta^M \), respectively.

Obviously, the eigenstates of \( \mathcal{H}_M \) are not identical to their vacuum counterparts. The splitting between the squared masses of the matter eigenstates is not the same as the vacuum splitting \( \Delta m^2 \), and the same happens with the mixing angle, the mixing in matter—the angle that rotates from the \( \nu_e, \nu_\mu \) basis, to the mass basis—is different from the vacuum mixing angle \( \theta \). Clearly however, all the physics of neutrino propagation in matter is controlled by the matter Hamiltonian \( \mathcal{H}_M \). However, according to Eq. (40), at least at the formal level, \( \mathcal{H}_M \) has the same functional dependence on the matter parameters \( \Delta m^2_M \) and \( \theta^M \) in as the vacuum Hamiltonian \( \mathcal{H}_{\text{Vac}} \), Eq. (28), on the vacuum ones, \( \Delta m^2 \) and \( \theta \). Therefore, \( \Delta m^2_M \) corresponds to the effective splitting between the squared masses of the eigenstates in matter, and \( \theta^M \) corresponds to the effective mixing angle in matter.

In a typical experimental set-up where the neutrino beam is generated by an accelerator and sent away to a detector that is, say, several hundred, or even thousand kilometers away, it traverses through earth matter, but only superficially, it does not get deep into the earth. The matter density met by such a beam en voyage can be taken to be approximately constant. Therefore, the electron density \( N_e \) is also constant, and the same happens with the parameter \( A \), and the matter Hamiltonian \( \mathcal{H}_M \). They all become approximately position independent, and therefore quite analogue to the vacuum Hamiltonian \( \mathcal{H}_{\text{Vac}} \), which was absolutely position independent. Comparing Eqs. (40) and (28), we can immediately conclude that since \( \mathcal{H}_{\text{Vac}} \) gives rise to the vacuum oscillation probability \( P(\nu_e \rightarrow \nu_\mu) \) of Eq. (33), \( \mathcal{H}_M \) must give rise to a matter oscillation probability of the form
\[ P_M(\nu_e \rightarrow \nu_\mu) = \sin^2 2\theta^M \sin^2 \left( \Delta m^2_M \frac{L}{4E} \right) \]

That is, the oscillation probability in matter (formally) is the same as in vacuum, except that the vacuum parameters \( \theta \) and \( \Delta m^2 \) are replaced by their matter counterparts.

In theory, judging simply its potential, matter effects can have very drastic effects. From Eq. (39) for the effective mixing angle in matter, \( \theta^M \), we can appreciate that even when the vacuum mixing angle \( \theta \) is incredible small, say, \( \sin^2 2\theta = 10^{-4} \), if we get to have \( A \cong \cos 2\theta \), then \( \sin^2 2\theta^M \) can be brutally enhanced as compared to its vacuum value and can even reach its maximum possible value, one. This brutal enhancement of a tiny mixing angle in vacuum up to a sizeable one in matter is the “resonant” version of the Mikheyev-Smirnov-Wolfenstein effect [6–9]. In the beginning of solar neutrino experiments, people entertained the idea that this brutal enhancement was actually taking place inside the sun. Nonetheless, as we will see soon the solar neutrino mixing angle is quite sizeable (≈ 34°) already in vacuum [10]. Then, although matter effects on the sun are important and they do enhance the solar mixing angle, unfortunately they are not as drastic as we once dreamt.
4 Evidence for neutrino oscillations

4.1 Atmospheric and accelerator neutrinos

Almost fifteen years have elapsed since we were presented convincing evidence of neutrino masses and mixings, and since then, the evidence has only grown. SuperKamiokande (SK) was the first experiment to present compelling evidence of νµ disappearance in their atmospheric neutrino fluxes, see [12]. In Fig. 2 the zenith angle (the angle subtended with the horizontal) dependence of the multi-GeV νµ sample is shown together with the disappearance as a function of L/E plot. These data fit amazingly well the naive two component neutrino hypothesis with

$$\Delta m_{\text{atm}}^2 = 2 - 3 \times 10^{-3} \text{eV}^2 \quad \text{and} \quad \sin^2 \theta_{\text{atm}} = 0.50 \pm 0.15$$

(42)

Roughly speaking SK corresponds to an L/E for oscillations of 500 km/GeV and almost maximal mixing (the mass eigenstates are nearly even admixtures of muon and tau neutrinos). No signal of an involvement of the third flavour, νe, is found so the assumption is that atmospheric neutrino disappearance is basically $\nu_\mu \longrightarrow \nu_\tau$.

Fig. 2: Superkamiokande’s evidence for neutrino oscillations both in the zenith angle and L/E plots

After atmospheric neutrino oscillations were established, two new experiments were built, sending (man-made) beams of νµ neutrinos to detectors located at large distances: the K2K experiment [13, 14], sends neutrinos from the KEK accelerator complex to the old SK mine, with a baseline of 120 km while the MINOS experiment [15], sends its beam from Fermilab, near Chicago, to the Soudan mine in Minnesota, a baseline of 735 km. Both experiments have seen evidence for νµ disappearance consistent with the one found by SK. The results of both are summarised in Fig. 3.

4.2 Reactor and solar neutrinos

The KamLAND reactor experiment, an antineutrino disappearance experiment, receiving neutrinos from sixteen different reactors, at distances ranging from hundred to thousand kilometers, with an average baseline of 180 km and neutrinos of a few ev, [16, 17], has seen evidence of neutrino oscillations. Such evidence was collected not only at a different L/E than the atmospheric and accelerator experiments but also consists on oscillations involving electron neutrinos, νe, the ones which were not involved before. These oscillations have also been seen for neutrinos coming from the sun (the sun produces only electron neutrinos). However, in order to compare the two experiments we should assume that neutrinos (solar)
and antineutrinos (reactor) behave in the same way, i.e. assume CPT conservation. The best fit values in the two neutrino scenario for the KamLAND experiment are

$$\Delta m^2 = 8.0 \pm 0.4 \times 10^{-5} \text{eV}^2 \quad \text{and} \quad \sin^2 \theta = 0.31 \pm 0.03$$

(43)

In this case, the $L/E$ involved is 15 km/MeV which is more than an order of magnitude larger than the atmospheric scale and the mixing angle, although large, is clearly not maximal.

Fig. 4 shows the disappearance probability for the $\bar{\nu}_e$ for KamLAND as well as several older reactor experiments with shorter baselines. The second panel depicts the flavour content of the $^8$Boron solar neutrino flux (with GeV energies) measured by SNO, [18], and SK, [19]. The reactor outcome can be explained in terms of two flavour oscillations in vacuum, given that the fit to the disappearance probability, is appropriately averaged over $E$ and $L$.

![Fig. 3](image_url)

**Fig. 3**: Allowed regions in the $\Delta m^2_{\text{atm}}$ vs $\sin^2 \theta_{\text{atm}}$ plane for MINOS data as well as for K2K data and two of the SK analyses. MINOS’s best fit point is at $\sin^2 \theta_{\text{atm}} = 1$ and $\Delta m^2_{\text{atm}} = 2.7 \times 10^{-3} \text{eV}^2$.

![Fig. 4](image_url)

**Fig. 4**: Disappearance of the $\bar{\nu}_e$ observed by reactor experiments as a function of distance from the reactor. The flavour content of the $^8$Boron solar neutrinos for the various reactions for SNO and SK. CC: $\nu_e + d \rightarrow e^- + p + p$, NC: $\nu_x + d \rightarrow \nu_x + p + n$ and ES: $\nu_\alpha + e^- \rightarrow \nu_\alpha + e^-$.

The analysis of neutrinos coming from the sun is slightly more sophisticated because it should include the matter effects that the neutrinos suffer since they are born (at the centre of the sun) until
they leave it, which are important at least for the $^8$Boron neutrinos. The pp and $^7$Be neutrinos are less energetic and therefore are not significantly altered by the presence of matter and leave the sun as if it were ethereal. $^8$Boron neutrinos on the other hand, leave the sun strongly affected by the presence of matter and this is evidenced by the fact that they leave the sun as the $\nu_2$ mass eigenstate and therefore do not undergo oscillations. This difference is, as mentioned, due mainly to their differences at birth. While pp ($^7$Be) neutrinos are created with an average energy of 0.2 MeV (0.9 MeV), $^8$B are born with 10 MeV and as we have seen the impact of matter effects grows with the energy of the neutrino.

However, we should stress that we do not really see solar neutrino oscillations. To trace the oscillation pattern, we need a kinematic phase of order one. In the case of neutrinos coming from the sun the kinematic phase is

$$\Delta_\odot = \frac{\Delta m^2_2 L}{4E} = 10^{7\pm1}. \quad (44)$$

Therefore, solar neutrinos behave as "effectively incoherent" mass eigenstates once they leave the sun, and remains being so once they reach the earth. Consequently the $\nu_e$ survival probability is given by

$$\langle P_{ee} \rangle = f_1 \cos^2 \theta_\odot + f_2 \sin^2 \theta_\odot \quad (45)$$

where $f_1$ is the $\nu_1$ content or fraction of $\nu_\mu$ and $f_2$ is the $\nu_2$ content of $\nu_\mu$ and therefore both fractions satisfy

$$f_1 + f_2 = 1. \quad (46)$$

However, as we have mentioned, pp and $^7$Be solar neutrinos are not affected by the solar matter and oscillate as in vacuum and thus, in their case $f_1 \approx \cos^2 \theta_\odot = 0.69$ and $f_2 \approx \sin^2 \theta_\odot = 0.31$. In the $^8$B a neutrino case, however, matter effects are important and the corresponding fractions are substantially altered, see Fig. 5.

Fig. 5: The sun produces $\nu_e$ in the core but once they exit the sun thinking about them in the mass eigenstate basis is useful. The fraction of $\nu_1$ and $\nu_2$ is energy dependent above 1 MeV and has a dramatic effect on the $^8$Boron solar neutrinos, as first observed by Davis.

In a two neutrino scenario, the day-time CC/NC measured by SNO, which is roughly identical to the day-time average $\nu_e$ survival probability, $\langle P_{ee} \rangle$, reads

$$\frac{CC}{NC} \bigg|_{\text{day}} = \langle P_{ee} \rangle = f_1 \cos^2 \theta_\odot + f_2 \sin^2 \theta_\odot, \quad (47)$$

where $f_1$ and $f_2 = 1 - f_1$ are the $\nu_1$ and $\nu_2$ contents of the muon neutrino, respectively, averaged over the $^8$B neutrino energy spectrum appropriately weighted with the charged current current cross section.
Therefore, the $\nu_1$ fraction (or how much $f_2$ differs from 100% ) is given by

$$f_1 = \left( \frac{CC}{NC} \right)_{day} - \sin^2 \theta_{\odot} \approx 10\%$$

where the central values of the last SNO analysis, [18], were used. As there are strong correlations between the uncertainties of the CC/NC ratio and $\sin^2 \theta_{\odot}$ it is not obvious how to estimate the uncertainty on $f_1$ from their analysis. Note, that if the fraction of $\nu_2$ were 100%, then $\frac{CC}{NC} \left|_{day} \right. = \sin^2 \theta_{\odot}$.

Using the analytical analysis of the Mikheyev-Smirnov-Wolfenstein (MSW) effect, provided in [20], one can obtain the mass eigenstate fractions, which are given by

$$f_2 = 1 - f_1 = (\sin^2 \theta_{\odot}^M + P_x \cos 2\theta_{\odot}^M)_{8B},$$

with $\theta_{\odot}^M$ being the mixing angle as given at the $\nu_e$ production point and $P_x$ is the probability of the neutrino to hop from one mass eigenstate to the second one during the Mikheyev-Smirnov resonance crossing. The average $\langle \ldots \rangle_{8B}$ is over the electron density of the $^8B$ $\nu_e$ production region in the centre of the Sun as given by the Solar Standard Model and the energy spectrum of $^8B$ neutrinos appropriately weighted with SNO’s charged current cross section. All in all, the $^8B$ energy weighted average content of $\nu_2$’s measured by SNO is

$$f_2 = 91 \pm 2\%$$ at the 95 % C.L. (50)

Therefore, it is obvious that the $^8B$ solar neutrinos are the purest mass eigenstate neutrino beam known so far and SK super famous picture of the sun taken (underground) with neutrinos is made with approximately 90% of $\nu_2$.

On March 8, 2012 a newly built reactor neutrino experiment, the Daya Bay experiment, located in China, announced the measurement of the third mixing angle [11], the only one which was still missing and found it to be

$$\sin^2(2\theta_{12}) = 0.092 \pm 0.017$$

The fact that this angle, although smaller that the other two, is still sizeable opens the door to a new generation of neutrino experiments aiming to answer the open questions in the field.

5 $\nu$ Standard Model

Now that we have understood the physics behind neutrinos oscillations and have learnt the experimental evidence about the parameters driving this oscillations, we can move ahead and construct the Neutrino Standard Model:

- it consists of three light ($m_i < 1$ eV) neutrinos, i.e. it involves only two mass differences $\Delta m^2_{atm} \approx 2.5 \times 10^{-3}$eV$^2$ and $\Delta m^2_{solar} \approx 8.0 \times 10^{-5}$eV$^2$.
- so far we have not seen any experimental indication (or need) for additional neutrinos. As we have measured long time ago the invisible width of the $Z$ boson and found it to be 3, if new neutrinos are going to be incorporated into the model, they cannot couple to the $Z$ boson, they cannot enjoy weak interactions, so we call them sterile. However, as sterile neutrinos have not been seen, and are not needed, our Neutrino Standard Model will contain only the three active flavours: $e$, $\mu$ and $\tau$.
- the unitary mixing matrix, called the PMNS matrix, which describes the relation between flavour eigenstates and mass eigenstates, comprises three mixing angles (the so called solar mixing angle:$\theta_{12}$,
2 Neutrino mass and character

6.1 Absolute neutrino mass

The absolute mass scale of the neutrino cannot be obtained in oscillation experiments, however this does not mean we cannot have it. Direct experiments like tritium beta decay, or neutrinoless double beta decay.
decay and indirect ones, like cosmological observations, have potential to feed us the information on the absolute scale of neutrino mass, we so desperately need. The Katrin tritium beta decay experiment, [22], has sensitivity down to 200 meV for the "mass" of $\nu_e$ defined as

$$m_{\nu_e} = |U_{e1}|^2 m_1 + |U_{e2}|^2 m_2 + |U_{e3}|^2 m_3.$$  \hspace{1cm} (52)

Fig. 7: The effective mass measured in double $\beta$ decay, in cosmology and in Tritium $\beta$ decay versus the mass of the lightest neutrino. Below the dashed lines, only the normal hierarchy is allowed.

Neutrino-less double beta decay experiments, see [23] for a review, do not measure the absolute mass of the neutrino directly but a combination of neutrino masses and mixings,

$$m_{\beta\beta} = |\sum_i m_i U_{ei}^2| = |m_1 c_{13}^2 c_{12}^2 + m_2 s_{13}^2 s_{12} e^{-2i\alpha} + m_3 s_{13}^2 e^{2i\beta}|,$$  \hspace{1cm} (53)

where it is understood that neutrinos are taken to be Majorana particles. The new generation of experiments seeks to reach below 10 meV for $m_{\beta\beta}$ in double beta decay.

Cosmological probes measure the sum of the neutrino masses,

$$m_{\text{cosmo}} = \sum_i m_i.$$  \hspace{1cm} (54)

If $\sum m_i \approx 50$ eV, the energy balance of the universe saturates the bound coming from its critical density. The current limit, [24], is a few % of this number, $\sim 1$ eV. These bounds are model dependent but
they do all give numbers of the same order of magnitude. However, given the systematic uncertainties characteristic of cosmology, a solid limit of less that 100 meV seems way too aggressive.

Fig. 7 shows the allowed parameter space for the neutrino masses (as a function of the absolute scale) for both the normal and inverted hierarchy.

### 6.2 Majorana vs Dirac

A fermion mass is nothing but a coupling between a left handed state and a right handed one. Thus, if we examine a massive fermion at rest, then one can regards this state as a linear combination of two massless particles, one right handed and one left handed particle. If the particle we are examining is electrically charged, like an electron, both particles, the left handed as well as the right handed must have the same charge (we want the mass term to be electrically neutral). This is a Dirac mass term. However, for a neutral particle, like a sterile neutrino, a new possibility opens up, the left handed particle can be coupled to the right handed anti-particle, (a term which would have a net charge, if the fields are not absolutely and totally neutral) this is a Majorana mass term.

Thus a neutral particle does have two ways of getting a mass term, a la Dirac or a la Majorana, and in principle can have them both, as shown:

\[
\begin{array}{c}
\text{Left Chiral} \\
\nu_L \\
\exists \\
\nu_R
\end{array}
\quad \leftrightarrow \quad 
\begin{array}{c}
\text{Dirac Masses} \\
\bar{\nu}_R \\
\exists \\
\bar{\nu}_L
\end{array}
\]

\[
\begin{array}{c}
\text{Majorana} \\
\text{Masses}
\end{array}
\]

In the case of a neutrino, the left chiral field couples to \(SU(2) \times U(1)\) implying that a Majorana mass term is forbidden by gauge symmetry. However, the right chiral field carries no quantum numbers, is totally and absolutely neutral. Then, the Majorana mass term is unprotected by any symmetry and it is expected to be very large, of the order of the largest scale in the theory. On the other hand, Dirac mass terms are expected to be of the order of the electroweak scale times a Yukawa coupling, giving a mass of the order of magnitude of the charged lepton or quark masses. Putting all the pieces together, the mass matrix for the neutrinos results as in Fig. 8.

\[
\begin{pmatrix}
0 & m_D \\
-m_D & M
\end{pmatrix}
\]

**Fig. 8:** The neutrino mass matrix with the various right to left couplings, \(M_D\) is the Dirac mass terms while 0 and \(M\) are Majorana masses for the charged and uncharged (under \(SU(2) \times U(1)\)) chiral components.

To get the mass eigenstates we need to diagonalise the neutrino mass matrix. By doing so, one is left with two Majorana neutrinos, one super-heavy Majorana neutrino with mass \(\sim M\) and one light Majorana neutrino with mass \(m_D^2/M\), i.e. one mass goes up while the other sinks, this is what we call the seesaw mechanism, [25–27]. The light neutrino(s) is(are) the one(s) observed in current experiments (its mass differences) while the heavy neutrino(s) are not accessible to current experiments and could be responsible for explaining the baryon asymmetry of the universe through the generation of a lepton
asymmetry at very high energy scales since its decays can in principle be CP violating (they depend on the two Majorana phases on the PNMS matrix which are invisible for oscillations).

If neutrinos are Majorana particles lepton number is no longer a good quantum number and a plethora of new processes forbidden by lepton number conservation can take place, it is not only neutrino-less double beta decay. For example, a muon neutrino can produce a positively charged muon. However, this process and any processes of this kind, would be suppressed by \((m_\nu/E)^2\) which is tiny, \(10^{-20}\), and therefore, although they are technically allowed, are experimentally unobservable.

7 Conclusions

The experimental observations of neutrino oscillations, meaning that neutrinos have mass and mix, answered questions that had endured since the establishment of the Standard Model. As those veils have disappeared, new questions open up and challenge our understanding:

– what is the nature of the neutrino? are they Majorana or Dirac? are neutrinos totally neutral?
– is the spectrum normal or inverted?
– is CP violated (is \(\sin \delta \neq 0\))?
– which is the absolute mass scale of the neutrinos?
– are there new interactions?
– can neutrinos violate CPT [28]?
– are these intriguing signals in short baseline reactor neutrino experiments (the missing fluxes) a real effect? Do they imply the existence of sterile neutrinos?

We would like to answer these questions. For doing it, we are doing right now, and we plan to do new experiments. These experiments will, for sure bring some answers and clearly open new, pressing questions. Only one thing is clear. Our journey into the neutrino world is just beginning.

Acknowledgements

I would like to thank the students and the organisers of the European School on HEP for giving me the opportunity to present these lectures in such a wonderful atmosphere. I did enjoy each day of the school enormously.

References


LHC Results - Highlights

G. Rolandi
CERN, Geneva, Switzerland and Scuola Normale Superiore, Pisa, Italy

Abstract
The LHC has delivered already 10 fb\(^{-1}\) of proton-proton collisions at a centre-of-mass energy of 7-8 TeV. With this data set, ATLAS and CMS have discovered a new boson at a mass of about 125 GeV and have searched for new physics at the TeV scale.

1 Introduction
The LHC [1] performs well above expectations with a peak luminosity of \(4 \times 10^{33}\) cm\(^{-2}\) s\(^{-1}\) from collisions of two 3.5 TeV proton beams in 2011 and two 4 TeV beams in 2012. In 2011, the LHC delivered about 6 fb\(^{-1}\) to ATLAS [2] and CMS [3]. With this large integrated luminosity, it is possible to search effectively for the Standard Model (SM) Higgs boson and to probe the existence of new physics at the TeV scale. This luminosity has been reached with an inter-bunch spacing of 50 ns and more than 1.2 \(10^{11}\) protons per bunch. At these currents, the average number of inelastic interactions per bunch crossing (pile-up) is about 15, hence posing new challenges to trigger, event reconstruction, and quality of reconstructed physics objects.

New algorithms have been designed to mitigate the effect of the pileup. In spite of the difficult experimental conditions, ATLAS and CMS have been able to calibrate quickly their data and to deliver new physics results shortly after the start of data taking. They are producing physics papers at a rate of about 100 papers per year per experiment, probing the Standard Model and searching for new physics. The main physics messages of the analyses of the data collected up to now are:

- the Standard Model is still in excellent shape;
- a new boson [4, 5] has been found with properties compatible with those predicted for the SM Higgs Boson;
- no sign of new physics has been found yet.

In this lecture, I will concentrate on two topics: search for Supersymmetry and search for the Higgs boson.

2 Supersymmetry
Supersymmetry (SUSY) [6–13] is a well-motivated extension of the SM. It introduces a large number of new particles with the same quantum numbers as their SM partners, but differing by half a unit of spin.

With R-parity conservation [14], the supersymmetric particles, such as squarks and gluinos, are produced in pairs and decay to the lightest, stable supersymmetric particle (LSP). If the LSP is neutral and weakly interacting, a typical signature is a final state of multi-jets and possibly leptons accompanied by large Missing Transverse Energy (MET).

The cross sections for producing SUSY particles are shown in Fig. 1 as a function of the mass of the particles. In quark and gluons collisions it is easy to produce coloured objects like gluinos and squarks, which decay typically to jets and MET, while the cross sections for Electroweak productions are smaller and the mass reach substantially reduced. These "ewkinos" decays typically produce many leptons and MET.

Both ATLAS [18–22] and CMS [23–27] presented many hadronic SUSY searches on the 7 TeV data based directly or indirectly on MET. These searches can be interpreted in many ways.
versions of SUSY, with a drastic reduction of the more than 100 parameter space like CMSSM [28] or mSugra [29], are excluded for gluinos and squarks below about 1 TeV and are now cornered. The searches can also be interpreted in terms of simplified models [30] where a single decay chain is considered with the assumption that the branching fractions along this chain are 100%.

Figure 2 shows two examples. The models assumed here are i) gluino pair production when squarks are much heavier than gluinos and the gluinos decay to two light quarks and a neutralino (left) and ii) squark-gluino associated production with gluino decaying into a quark pair and a neutralino and the squark decaying into quark neutralino. The neutralino here is assumed to be massless (right).

While generic SUSY production at the scale of about 1 TeV is not compatible with data, there is still room for natural models [31, 32] where gluinos and third-generation squarks are below 1 TeV. Both ATLAS and CMS have done specific searches for the third-generation squarks. Examples are the search for same-sign lepton pairs, b jets and MET by CMS [33] - motivated by final states including four top or two top and two W as shown in Fig. 3 (a) and (b) - and the search for three b-jets and MET by ATLAS [34] addressing models shown in Fig 3 (c) and (d). Broadly speaking, these searches exclude gluinos of 900 GeV for third generation squarks lighter than 300–400 GeV.

Fig. 1: Cross sections for producing SUSY particles in pp collisions at $\sqrt{s} = 7$ TeV computed with PROSPINO [15–17].

Fig. 2: Upper limit [27] on the cross section for gluino pair production when squarks are much heavier than gluinos and the gluinos decay to two light quarks and a neutralino (left). Upper limit [22] on squark-gluino associated production with the gluino decaying into a quark pair and a neutralino and the squark decaying into quark neutralino. The neutralino here is assumed to be massless (right).
Both ATLAS and CMS have also performed a model-independent search for Weakly Interacting Massive Particles (WIMP) production [35, 36] triggering on events with a monojet and MET. Here the jet is produced by initial state radiation of one of the interacting partons and the two WIMPs escape undetected leading to spectacular events like the one shown in Fig. 4: the detector is empty with the exception of a single high energy jet. In the SM these events are produced by a high transverse momentum Z decaying into a neutrino pair. This background can be effectively measured from similar events where the Z decays into a muon pair.

![Monojet event recorded by CMS](image)

The measured cross section is compatible with the SM background and limits can be set on generic WIMP production. For dark matter models, the observed limit on the cross section depends on the mass of the dark matter particle and the nature of its interaction with the SM particles. The limits on the effective contact interaction scale as a function of the wimp mass can be translated into a limit on the dark matter-nucleon scattering cross section [37]. These limits can be compared with the constraints from direct and indirect detection experiments. The LHC limits [36] are competitive with those from direct WIMP search for the spin-dependent interaction for WIMP mass below few hundreds GeV, and also for the spin-independent interaction for WIMP mass below 10 GeV.

### 3 Search for the Higgs boson

Three weeks after I gave this lecture in Anjoux, ATLAS [4] and CMS [5] announced the observation of a new boson compatible with the SM Higgs boson in a CERN seminar. In this section, with the agreement of the organizers of the school, I describe the observation of the boson.

The search for the Higgs boson (H) and the justification of the electroweak symmetry breaking [38, 39] was one of the main reasons for the construction of the Large Hadron Collider. This search was therefore a high priority analysis for ATLAS and CMS. In the SM the cross section for H production in proton proton collision at 7 TeV is about 17 pb [40] for $m_H \simeq 125$ GeV and 30% higher at 8 TeV. In about 10 fb$^{-1}$ of integrated luminosity, shared about equally between the two energies, some 200,000 Higgs bosons are produced in each experiment. It is very difficult however to separate this signal from
the very large background of SM processes, especially the hadronic final states, and specific searches
with leptons or photons in the final state are performed. Table 1 lists the search channels together with
the branching ratio expected for a SM Higgs boson.

A value of $m_H \simeq 125$ GeV is smaller than the sum of the masses of the vector boson (V) pairs,
in the decays $H \to V$ one or both V are off mass-shell. In $H \to 4\ell$ and $H \to \gamma\gamma$ the H mass
is reconstructed with high resolution (1-2%) with the precise measurements of the momenta of leptons and
photons. Because the width of the SM Higgs boson for $m_H \simeq 125$ GeV is a few MeV, one expects
to see a narrow peak dominated by the instrumental resolution. The other channels have worse mass
resolution because of the missing neutrinos in the W and $\tau$ decays or because the b-jets are reconstructed
with some 10% resolution. The search in the bb final state is performed in the associate production
$VH \to Vbb$ where the V decays into leptons that provide the trigger and reduce the overwhelming
hadronic background.

Figure 5 displays the expected discovery potential for a SM Higgs boson in the CMS experiment
as a function of $m_H$ (ATLAS have similar figures). The probability that the background can produce a
fluctuation greater than the potential excess produced in data by a SM Higgs boson (the so called local
p-value) is estimated less than $10^{-8}$ i.e. more than five standard deviations. The most sensitive channels
are those where a narrow peak can be observed: $H \to 4\ell$ and $H \to \gamma\gamma$. The WW channel is also quite
sensitive.

### 3.1 $H \to WW \to \ell\nu \ell'\nu'$

The decay mode $H \to WW \to \ell\nu \ell'\nu'$ is the main search channel for a SM Higgs boson with mass
above the WW threshold of 160 GeV. With good experimental control of the MET and very tight lepton

### Table 1: Channels used in the search for the Higgs boson and the branching ratios expected in the SM for $m_H = 125$ GeV. Here $\ell$ and $\ell'$ indicate an electron or a muon.

<table>
<thead>
<tr>
<th>Channel</th>
<th>Branching Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H \to ZZ \to \ell\ell \ell'\ell'$</td>
<td>$1.2 \times 10^{-4}$</td>
</tr>
<tr>
<td>$H \to \gamma\gamma$</td>
<td>$2.3 \times 10^{-3}$</td>
</tr>
<tr>
<td>$H \to WW \to \ell\nu \ell'\nu'$</td>
<td>$1.0 \times 10^{-2}$</td>
</tr>
<tr>
<td>$H \to \tau\tau$</td>
<td>$6.0 \times 10^{-2}$</td>
</tr>
<tr>
<td>$H \to bb$</td>
<td>$5.8 \times 10^{-1}$</td>
</tr>
</tbody>
</table>
identification, it is possible to reject large part of the reducible background and extend the sensitivity down to $m_H \approx 120$ GeV. The signature is two isolated opposite-sign charged leptons and large MET caused by the two undetected neutrinos. The most sensitive channel is when the two leptons have opposite flavour and there are no extra jets in the event. Here the main background is the irreducible non resonant WW production and the reducible W+jet production when the jet fakes a lepton. The other channels with same flavour leptons or with associated jets have larger backgrounds from Drell-Yan and top quark decay respectively and contribute less than 20% to the sensitivity. The yields of the largest backgrounds are estimated from control regions. One important variable to separate the signal from the irreducible background is the angle between the two leptons. Due to spin correlations, this variable is small for W pairs from the spin-0 H decay and large for the WW non resonant production [41]. In ATLAS the transverse mass of the MET vector and the di-lepton system - shown in Fig. 6 (a) - is used to test for the presence of a signal for all jet multiplicities. In CMS, the signal is separated from the background with kinematical and topological requirements optimized for each mass hypothesis; one of the most sensitive variables is the dilepton invariant mass shown in Fig. 6 (b).

![Graphs showing H → WW analysis](a) ![Graphs showing dilepton invariant mass](b)

**Fig. 6:** H → WW analysis. a) ATLAS, Distribution of the transverse mass, in the 0-jet and 1-jet analyses with opposite flavour, for events satisfying all selection criteria. The expected signal for $m_H = 125$ GeV is shown added to the background prediction. The hashed area indicates the total uncertainty on the background prediction. b) CMS, Distribution of dilepton invariant mass for the zero-jet opposite flavour category at 8 TeV after the full selection, except for the selection on $m_{\ell\ell}$ itself. The signal expected from a Higgs boson with a mass $m_H = 125$ GeV is shown added to the background.

In the most sensitive channel, leptons with opposite flavour and zero jets, CMS observe 158 events, estimate a background of $124 \pm 12$ events and expect $24 \pm 5$ events for a SM Higgs boson of 125 GeV. Similarly ATLAS observe 185 events, estimate a background of $142 \pm 16$ events and expect $20 \pm 4$ events for a SM Higgs boson of 125 GeV.

### 3.2 H → ZZ → ℓℓ ℓ′ ℓ′′

The decay mode $H \rightarrow ZZ \rightarrow \ell\ell\ell'\ell''$, the so called golden channel, is characterized by a small signal yield over a flat irreducible background of direct ZZ production. Because the signal yield is small, it is important to maximize the efficiency lowering as much as possible the threshold on the lepton transverse momenta: at $m_H = 125$ GeV the average transverse momentum of the softest lepton is about 7 GeV. Also the lepton identification is relaxed in order to maximize the efficiency. The reducible background is evaluated with control regions and is small, in spite of the relaxed identification criteria, thanks to the
presence of four leptons in the final state.

In the ATLAS analysis the thresholds for muons and electrons are 6 and 7 GeV. In CMS, they are 5 and 7 GeV. The events are selected pairing opposite charge and same flavour leptons and with requirements on the invariant masses of the pairs. A recovery of the final state radiation photons is also used in CMS. The expected background yield of the irreducible background is estimated using the MC simulation normalized to the theoretical cross section for ZZ production. The identification efficiency, the energy scale and the energy resolution are measured using large samples of Z, Y and J/ψ decaying into two leptons.

Figure 7 shows the invariant mass distribution of the selected events compared to the estimated background. The peak at about 90 GeV is the decay of the Z into four leptons, where a lepton pair is radiated by one of the leptons originating from the Z decay. In this process the distribution of the lowest momentum lepton is softer than in the Higgs boson decays. The Z peak has a different yield in the two experiments because of the larger efficiency in CMS for low momentum leptons. The reducible background in CMS is indeed much smaller than the reducible one, in ATLAS they are comparable for \( m_H = 125 \) GeV.

In the four-lepton mass region 121.5-130.5 GeV CMS observe 9 events, estimate a background of 3.8 ± 0.5 events and expect 7.5 ± 0.9 events for a SM Higgs boson of 125 GeV. In the four-lepton mass region 120-130 GeV ATLAS observe 13 events, estimate a background of 4.8 ± 0.2 events and expect 5.3 ± 0.5 events for a SM Higgs boson of 125 GeV.

The scalar nature of the Higgs boson provides important discriminating power between the signal and the irreducible background. The kinematics of the ZZ \( \rightarrow \ell\ell'\ell\ell' \) process is fully described by five angles and the invariant masses of the two lepton pairs [42–44] for a fixed invariant mass of the four-lepton system. In CMS, a kinematic discriminant (KD) is constructed based on the probability ratio of the signal and background hypotheses as described in Ref. [45]. Figure 8 shows the distribution of four leptons invariant mass versus KD for the selected events. The discriminant KD takes large values for signal like events and small values for background like events.
A clustering of events is observed with a high value of the kinematic discriminant at $m_H \simeq 125$ GeV where the background expectation is low, corresponding to the excess seen in the one-dimensional mass distribution.

### 3.3 $H \rightarrow \gamma\gamma$

The decay mode $H \rightarrow \gamma\gamma$ is characterized by a narrow peak in the diphoton invariant mass distribution above a large irreducible background from QCD production of two photons and a reducible background where one - and in few cases two - reconstructed photons originate from misidentification of jets. The resolution in the invariant mass of the two photons varies on event by event basis, depending on the properties of the reconstructed photons and of the overall event properties (e.g. number of reconstructed vertices). In order to improve the sensitivity of the analysis the events are categorized in exclusive sets with varying signal purity. The events with two reconstructed jets belong to separate categories to exploit the better signal to background ratio in the $H$ production via Vector Boson Fusion (VBF) where two jets originating from the two scattered quarks are expected with large difference in rapidity.

The diphoton invariant mass is reconstructed from the energies measured by the calorimeter and the position of the primary vertex. In ATLAS, the primary vertex of the hard interaction is identified exploiting the directions of flight of the photons as determined with the longitudinal segmentation of the electromagnetic calorimeter. The CMS calorimeter has no pointing information and the vertex is identified from the kinematic properties of the tracks associated with that vertex and their correlation with the diphoton kinematics.

In CMS, a multivariate regression algorithm is used to extract the photon energy and a photon-by-photon estimate of the uncertainty in that measurement. This information is used in a boosted decision tree (BDT) together with photon and vertex quality variables and kinematic variables. The BDT is trained to separate signal and background events and its output is used to assign the events in an optimal way to four categories with different signal-to-background ratio.

In ATLAS, the non-VBF events are separated in nine categories defined by the rapidity of the photons, the component of the diphoton transverse orthogonal to the axis defined by the difference between the two photon momenta [46,47] and the presence or not of converted photons.

The background in each category is estimated from data by fitting the measured diphoton mass spectrum with a background model with free shape and normalization. This background model is chosen...
Fig. 9: $H \rightarrow \gamma\gamma$. The distributions of the invariant mass of diphoton candidates after all selections are shown together with the result of a fit to the data of the sum of a signal component and a background component. (left) ATLAS. The panel in the middle (c) shows the weighted sample with the weights are explained in the text. The signal is fixed to $m_H = 126.5$ GeV and the background is described by a fourth-order Bernstein polynomial. The residuals of the weighted data with respect to the fitted background component are displayed in (d). The panel on the top (b) shows the residuals of the un-weighted data with respect to their fitted background. (right) CMS. The panel shows the weighted sample where the background is fitted with a fifth order polynomial. The coloured bands represent the $\pm 1$ and $\pm 2$ deviation uncertainties on the background estimate. The inset shows the central part of the unweighted invariant mass distribution by requiring that the potential residual bias is smaller than 20% of the statistical accuracy of the fit. The statistical analysis of the data is done with an unbinned likelihood function in each category.

The distribution of the invariant mass of the diphoton candidates summed an all categories is shown in Fig. 9. In order to exploit the different sensitivity of each category, the events are weighted with category-dependent factors reflecting the different signal-to-background ratios. In ATLAS, the weights are defined as $\ln(1 + S/B)$ where S is 90% of the expected signal and B is the integral of the background fit in the window containing S. In CMS, the weights proportional to $S/(S + B)$, where S and B are defined [almost] as in ATLAS and the weights are normalized such that the integral of the weighted signal model matches the number of signal events given by the best fit. An excess near 125 GeV appears clearly in both the weighted and unweighted distributions.

3.4 Statistical analysis

A common statistical procedure [48] for the interpretation of the SM Higgs boson searches has been developed by ATLAS and CMS. Data and background predictions are compared to the expected SM Higgs boson signal and the ratio $\mu$ between the measured and predicted signal strength is evaluated. The background-only hypothesis corresponds to $\mu = 0$ while a positive value of $\mu$ significantly different from zero indicates the presence of a signal. The signal expected for a SM Higgs boson is $\mu = 1$. At each mass, possible values of $\mu$ are tested with a test statistics based on the profile likelihood ratio [49] to extracts the information on the signal strength from a full likelihood fit to the data. The likelihood function includes all the parameters describing the data and all the parameters that describe the systematic uncertainties and their correlations.
The interpretation strategy is based on the modified frequentist criterion [50, 51] CLs. A value of $\mu$ is excluded at 95% CL when CLs is less than 5%. For each mass, the value of $\mu$ excluded at 95% CL, $\mu_{95}$, is computed. In practice a scan in steps of a fraction of the mass resolution is done for each channel. Figure 10 shows the value of $\mu_{95}$ as a function of the Higgs boson mass for the $\gamma\gamma$ and $4\ell$ channels.

Fig. 10: Top: $\gamma\gamma$ channel. Bottom: $4\ell$ channel. Left: ATLAS. Right: CMS. The expected exclusion limit $\mu_{95}$ computed in the background only hypothesis is shown together with its 1$\sigma$ (green) and 2$\sigma$ (yellow) bands as a function of the Higgs boson mass. The black line shows the observed $\mu_{95}$.

In this plot, nearby mass points are correlated with a correlation length that is given by the mass resolution of about 1-2%. The data follow roughly the expected value of $\mu_{95}$ with the exception of the region at $m_H \simeq 125$ GeV where a peak exceeding the 2$\sigma$ band is observed in all distributions.

The significance of the excess is quantified by the probability for a background fluctuation to be at least as large as the observed excess. This local p-value is shown in Fig. 11 for the combination of channels presented by ATLAS and CMS. The combination assumes the SM branching fractions.

At $m_H \simeq 125$ GeV ATLAS expect 5$\sigma$ and observe 6$\sigma$ while CMS expect 6$\sigma$ and observe 5$\sigma$. A clear signal is established in each experiment. The combination of the two experiments has a global significance in large excess of the 5$\sigma$ value that is the value conventionally required for claiming an observation. Since this new resonance decays to two photons, it must be a boson with spin different from 1 [52, 53]. Its mass can be measured by fitting the signal strength as a function of the mass in the most sensitive channels: $\gamma\gamma$ and $4\ell$. This fit allows the signal strength in each channel to float independently in order to reduce the model dependence. As a result, CMS quotes a mass of $125.3 \pm 0.4 \pm 0.5$ GeV and
ATLAS quotes a mass of 126.0 ± 0.4 ± 0.4 GeV. In both experiments, the main systematic error comes from the energy scales of electrons and photons evaluated from a comparison of data and simulation at the Z peak.

The best-fit signal strength is fit for each search channel independently at the measured value of the mass of the boson. The results of these fits are shown in Fig. 12. There is large consistency between the measurements of the two experiments. The bb and ττ channels, with low sensitivity, are compatible with µ = 0 and with µ = 1. The WW and ZZ channels have strengths close to µ = 1 in both experiments, with ATLAS consistently larger than CMS. The strength of the γγ channel is larger than 1 in both experiments: ATLAS measures µ = 1.8 ± 0.5 and CMS µ = 1.6 ± 0.4. The average of the two values gives µ = 1.7 ± 0.3, some two sigmas above the SM expectation.

An important quantity is the best fit value of the signal strength µ as a function of the Higgs boson mass for the combination of all search channels in each experiment. The combination assumes the SM branching fractions. This quantity is shown in Fig. 13. The signal strength is compatible with the SM expectation: ATLAS measures a slight excess µ = 1.4 ± 0.3 and CMS measures µ = 0.87 ± 0.23. The average gives µ = 1.1 ± 0.2.
Fig. 13: Best-fit signal strength as a function of the Higgs boson mass hypothesis for the full combination of the 2011 and 2012 data. The band shows the ±1σ uncertainty. Left: ATLAS. Right: CMS.

4 Conclusions

The Standard Model has passed the first scrutiny by LHC, which probed for the first time the TeV scale with data collected in 2011. The Higgs particle was the main missing block of the SM. The new boson found by ATLAS and CMS in the range of masses preferred by the precision electroweak tests is a spectacular confirmation of the SM framework. Still the SM leaves too many open questions to be considered a complete description of Nature.

Supersymmetry, considered as one of the most natural extensions of the SM, has been tested already with 2011 data. Direct searches exclude constrained SUSY models. The room for natural supersymmetry is quite restricted but scenarios with light stops or sbottoms are still open.

We look forward to the analysis of the new data collected by LHC in 2012 and to the higher energy run that will start after the energy upgrade to 13.5 TeV in 2015.

References


[22] ATLAS Collaboration, G. Aad et al., Search for squarks and gluinos with the ATLAS detector in final states with jets and missing transverse momentum using 4.7 fb-1 of 7 TeV proton-proton collision data, 1208.0949.


[34] ATLAS Collaboration, G. Aad et al., “Search for top and bottom squarks from gluino pair production in final states with missing transverse energy and at least three b-jets with the ATLAS detector,” 1207.4686.


LHC RESULTS–HIGHLIGHTS

Practical Statistics for Particle Physicists

H. B. Prosper
Florida State University, Tallahassee, USA

Abstract
We introduce a few of the key ideas of statistical analysis using two real-world examples to illustrate how these ideas are used in practice.

1 Introduction
These lectures introduce to two broad classes of theories of inference, the frequentist and Bayesian approaches. Two points should be made immediately. The first is that there is no such thing as “the” answer in statistics. Instead there are answers based on assumptions on which reasonable people may disagree. Second, none of the current theories of inference is perfect. It is worth appreciating these points in order to avoid fruitless arguments that cannot be resolved because they are ultimately about intellectual taste and not mathematical correctness.

For in-depth expositions of statistical analysis, we highly recommend the excellent books on statistics written for physicists, by physicists [1–4] and the very insightful book on the history of the ideas by Chatterjee [5].

2 Lecture 1: descriptive statistics, probability and likelihood
2.1 Descriptive statistics
Suppose we have a sample of $N$ data $X = x_1, x_2, \ldots, x_N$. It is often useful to summarize these data with a few numbers called statistics. A statistic is any number that can be calculated from the data and known parameters. For example, $t = (x_1 + x_N)/2$ is a statistic, but if the value of $\theta$ is unknown $t = (x_1 - \theta)^2$ is not. However, we particle physicists tend to refer to any function of the data as a statistic including those that contain unknown parameters.

The two most important statistics are

the sample mean (or average) \[
\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{1}
\]

and the sample variance \[
s^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2, \tag{2}
\]

\[= \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \bar{x}^2, \]

\[= \bar{x}^2 - \bar{x}^2. \]

The sample average is a measure of the center of the distribution of the data, while the sample variance is a measure of its spread. Statistics that merely characterize the data are called descriptive statistics, of which the sample average and variance are the most important.

Descriptive statistics can always be calculated because they depend only on a data sample $X$. We now consider numbers that cannot be calculated from the data alone. Imagine the repetition, infinitely many times, of the data generating system that yielded our data sample $X$, thereby creating an infinite set of data sets. We shall refer to the data generating system as an experiment and the infinite set of the results of the experiments as an infinite ensemble. This is clearly an abstraction.
The most common operation to perform on an ensemble is to compute the ensemble average of the statistics, which yield numbers such as the following.

<table>
<thead>
<tr>
<th>Ensemble average</th>
<th>$\langle x \rangle$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\mu$</td>
</tr>
<tr>
<td>Error</td>
<td>$\epsilon = x - \mu$</td>
</tr>
<tr>
<td>Bias</td>
<td>$b = \langle x \rangle - \mu$</td>
</tr>
<tr>
<td>Variance</td>
<td>$V = \langle (x - \langle x \rangle)^2 \rangle$</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>$\sigma = \sqrt{V}$</td>
</tr>
<tr>
<td>Mean square error</td>
<td>MSE = $\langle (x - \mu)^2 \rangle$</td>
</tr>
<tr>
<td>Root MSE</td>
<td>$\text{RMS} = \sqrt{\text{MSE}}$</td>
</tr>
</tbody>
</table>

None of these numbers can be calculated from data because the data needed do not objectively exist. Even in an experiment simulated on a computer, there are very few of these numbers we can calculate. If we know the mean $\mu$, perhaps because we have chosen its value, we can certainly calculate the error $\epsilon$ for any simulated datum $x$. But, we can only approximate the ensemble average $\langle x \rangle$, bias $b$, variance $V$, and MSE, since the ensembles available either on our computers or in the real world are always finite. The point is that the numbers that characterize the infinite ensemble are also abstractions.

The MSE is the most widely used measure of the closeness of an ensemble of numbers to some parameter $\mu$. The square root of the MSE is called the root mean square (RMS)$^1$. The MSE can be written as

$$\text{MSE} = V + b^2,$$

Exercise 1: Show this

the sum of the variance and the square of the bias, a very important result with practical consequences. For example, suppose that $\mu$ represents the mass of the Higgs boson and $x$ is a complicated function that is considered an estimator of the mass. An estimator is any function, which when data are entered into it, yields an estimate of the quantity of interest.

As noted, many of the numbers listed in Eq. (3) cannot be calculated because the information needed is unknown. This is true, in particular, of the bias. However, sometimes it is possible to relate the bias to another ensemble quantity. Consider the ensemble average of the sample variance, Eq. (2).

$$\langle s^2 \rangle = \langle x^2 \rangle - \langle \bar{x}^2 \rangle,$$

$$= V - \frac{V}{N},$$

Exercise 2a: Show this

Exercise 2b: Use the method Rndm() of the Root class TRandom3 to approximate the quantities in Eq. (3).

2.2 Probability

When the weather forecast specifies that there is a 80% chance of snow tomorrow at CERN, most people have an intuitive sense of what this means. Likewise, most people have an intuitive understanding of

---

$^1$Sometimes, the RMS and standard deviation are using interchangeably. However, the RMS is computed with respect to $\mu$, while the standard deviation is computed with respect to the ensemble average $\langle x \rangle$. The RMS and standard deviations are identical only if the bias is zero.
what it means to say that there is a 50-50 chance for a tossed coin to land heads up. Probabilistic
ideas are thousands of years old, but, starting in the sixteenth century these ideas were formalized into
increasingly rigorous mathematical theories of probability. In the theory formulated by Kolmogorov in
1933, Ω is some fixed mathematical space, \( E_1, E_2, \ldots \subseteq \Omega \) are subsets (called events) defined in some
reasonable way\(^2\), and \( P(E_j) \) is a number associated with subset \( E_j \). These numbers satisfy the

**Kolmogorov Axioms**

1. \( P(E_j) \geq 0 \)
2. \( P(E_1 + E_2 + \cdots) = P(E_1) + P(E_2) + \cdots \) for disjoint subsets
3. \( P(\Omega) = 1 \).

Consider two subsets \( A = E_1 \) and \( B = E_2 \). The quantity \( AB \) means \( A \) and \( B \), while \( A + B \) means
\( A \) or \( B \), with associated probabilities \( P(AB) \) and \( P(A + B) \), respectively. Kolmogorov assumed, not
unreasonably given the intuitive origins of probability, that probabilities sum to unity; hence the axiom
\( P(\Omega) = 1 \). However, this assumption can be dropped so that probabilities remain meaningful even if
\( P(\Omega) = \infty \) [6].

Figure 1 suggests another probability, namely, the number \( P(A|B) = P(AB)/P(B) \), called the
**conditional probability** of \( A \) given \( B \). This permits statements such as: “the probability that this track
was created by an electron given the measured track parameters” or “the probability to observe 17 events
given that the mean background is 3.8 events”. Conditional probability is a very powerful idea, but the
term itself is misleading. It implies that there are two kinds of probability: conditional and unconditional.
In fact, all probabilities are conditional in that they always depend on a specific set of conditions, namely,
those that define the space \( \Omega \). It is entirely possible to embed a family of subsets of \( \Omega \) into another space
\( \Omega' \) which assigns to each family member a different probability \( P' \). A probability is defined only relative
to some space of possibilities \( \Omega \).

\( A \) and \( B \) are said to be mutually exclusive if
\( P(AB) = 0 \), that is, if the truth of one denies the
truth of the other. They are said to be exhaustive if
\( P(A) + P(B) = 1 \). Figure 1 suggests the theorem

\[
P(A + B) = P(A) + P(B) - P(AB),
\]

**Exercise 3:** Prove theorem

which can be deduced from the rules given above. Another useful theorem is an immediate
consequence of the commutativity of “anding” \( P(AB) = P(BA) \) and the definition of \( P(A|B) \),
namely,

**Bayes Theorem**

\[
P(B|A) = \frac{P(A|B)P(B)}{P(A)},
\]

which provides a way to convert the probability
\( P(A|B) \) to the probability \( P(B|A) \). Using Bayes
theorem, we can, for example, deduce the probability \( P(e|x) \) that a particle is an electron, \( e \), given a set
of measurements, \( x \), from the probability \( P(x|e) \) of a set of measurements given that the particle is an
electron.

---

\(^2\)If \( E_1, E_2, \ldots \) are meaningful subsets of \( \Omega \), so to is the complement \( \overline{E}_1, \overline{E}_2, \ldots \) of each, as are countable unions and
intersections of these subsets.
2.2.1 Probability distributions

In this section, we illustrate the use of these rules to derive more complicated probabilities. First we start with a definition:

A Bernoulli trial, named after the Swiss mathematician Jacob Bernoulli (1654 – 1705), is an experiment with only two possible outcomes: \( S = \) success or \( F = \) failure.

Example

Each collision between protons at the Large Hadron Collider (LHC) is a Bernoulli trial in which something interesting happens (\( S \)) or does not (\( F \)). Let \( p \) be the probability of a success, which is assumed to be the same for each trial. Since \( S \) and \( F \) are exhaustive, the probability of a failure is \( 1 - p \). For a given order \( O \) of \( n \) proton-proton collisions and exactly \( k \) successes, and therefore exactly \( n - k \) failures, the probability \( P(k, O, n, p) \) is given by

\[
P(k, O, n, p) = p^k (1 - p)^{n-k}.
\]

If the order \( O \) of successes and failures is judged to be irrelevant, we can eliminate the order from the problem by summing over all possible orders,

\[
P(k, n, p) = \sum_O P(k, O, n, p) = \sum_O p^k (1 - p)^{n-k}.
\]

This procedure is called marginalization. It is one of the most important operations in probability calculations. Every term in the sum in Eq. (8) is identical and there are \( \binom{n}{k} \) of them. This yields the binomial distribution,

\[
\text{Binomial}(k, n, p) \equiv \binom{n}{k} p^k (1 - p)^{n-k}.
\]

By definition, the mean number of successes \( a \) is given by

\[
a = \sum_{k=0}^{n} k \text{ Binomial}(k, n, p),
= pn.
\]

Exercise 4: Show this

At the LHC \( n \) is a number in the trillions, while for successes of interest such as the creation of a Higgs boson the probability \( p \ll 1 \). In this case, it proves convenient to consider the limit \( p \to 0, n \to \infty \) in such a way that \( a \) remains constant. In this limit

\[
\text{Binomial}(k, n, p) \to e^{-a} a^k / k!,
\equiv \text{Poisson}(k, a).
\]

Exercise 5: Show this

Below we list the most common probability distributions.

Discrete distributions

\[
\text{Binomial}(k, n, p) \equiv \binom{n}{k} p^k (1 - p)^{n-k}
\]
Poisson\((k, a)\) 
\[
a^k \exp(-a) / k!
\]
Multinomial\((k, n, p)\) 
\[
\frac{n!}{k_1! \cdots k_K!} \prod_{i=1}^{K} p_i^{k_i}, \quad \sum_{i=1}^{K} p_i = 1, \sum_{i=1}^{K} k_i = n
\]

Continuous densities

Uniform\((x, a)\) 
\[
1/a
\]
Gaussian\((x, \mu, \sigma)\) 
\[
\exp\left[-\frac{(x - \mu)^2}{(2\sigma^2)}\right]/(\sigma\sqrt{2\pi})
\]
(also known as the Normal density)
LogNormal\((x, \mu, \sigma)\) 
\[
\exp\left[-\frac{(\ln x - \mu)^2}{(2\sigma^2)}\right]/(x\sigma\sqrt{2\pi})
\]
Chisq\((x, n)\) 
\[
x^{n/2-1} \exp(-x/2)/(2^{n/2}\Gamma(n/2))
\]
Gamma\((x, a, b)\) 
\[
x^{a-1}a^b \exp(-ax)/\Gamma(b)
\]
Exp\((x, a)\) 
\[
a^x/\Gamma(a)
\]
Beta\((x, n, m)\) 
\[
\frac{\Gamma(n+m)}{\Gamma(m)\Gamma(n)}x^{n-1}(1-x)^{m-1}
\]

Particle physicists tend to use the term probability distribution for both discrete and continuous functions, such as the Poisson and Gaussian distributions, respectively. But, strictly speaking, the continuous functions are probability densities, not probability distributions. In order to compute a probability from a density we need to integrate the density over a finite set in \(x\).

2.3 Likelihood

Let us assume that \(p(x|\theta)\) is a probability density function (pdf) such that \(P(A|\theta) = \int_A p(x|\theta) \, dx\) is the probability of the statement \(A = x \in R_x\), where \(x\) denotes possible data, \(\theta\) the parameters that characterize the probability model (that is the probability together with all the assumptions on which it is based), and \(R_x\) is a finite set. We shall use probability model as shorthand for probability density function (for continuous variables) or probability mass function (pmf) (basically, probabilities for discrete variables). If \(x\) is discrete, then both \(p(x|\theta)\) and \(P(A|\theta)\) are probabilities. The likelihood function is simply the probability model \(p(x|\theta)\) evaluated at the data \(x_O\) actually obtained, i.e., the function \(p(x_O|\theta)\). The following are examples of likelihoods.

**Example 1**

In 1995, CDF and DØ discovered the top quark [8, 9] at Fermilab. The DØ Collaboration found \(x = N = 17\) events. For a counting experiment, the datum can be modeled using
\[
p(x|n) = \text{Poisson}(x, n) \quad \text{probability to get} \ x \ \text{events}
\]
\[
p(N|n) = \text{Poisson}(N, n) \quad \text{likelihood of} \ N \ \text{events}
\]
\[
= n^N \exp(-n)/N!
\]
We shall analyze this example in detail in Lectures 2 and 3.

**Example 2**

Figure 2 shows a plot of the distance modulus versus redshift for \(N = 580\) Type 1a supernovae [7]. These heteroscedastic data\(^3\) \(D = \{z_i, x_i \pm \sigma_i\}\) are modeled using the likelihood
\[
p(D|\Omega_M, \Omega_L, Q) = \prod_{i=1}^{N} \text{Gaussian}(x_i, \mu_i, \sigma_i),
\]
\(^3\)Data in which each item, \(x_i\), or group of items has a different uncertainty.
which is an example of an un-binned likelihood. The cosmological model is encoded in the distance modulus function $\mu_i$, which depends on the redshift $z_i$ and the matter density and cosmological constant parameters $\Omega_M$ and $\Omega_\Lambda$, respectively. (See Ref. [10] for an accessible introduction to the analysis of these data.)

**Example 3**

The discovery of a Higgs boson by ATLAS [11] and CMS [12] in the di-photon final state ($pp \rightarrow H \rightarrow \gamma\gamma$) made use of an un-binned likelihood of the form,

$$p(x|s, m, w, b) = \exp[-(s + b)] \prod_{i=1}^{N} [s f_s(x_i|m, w) + b f_b(x_i)]$$

where $x = \text{di-photon masses}$
- $m = \text{mass of boson}$
- $w = \text{width of resonance}$
- $s = \text{expected (i.e., mean) signal count}$
- $b = \text{expected background count}$
- $f_s = \text{signal probability density}$
- $f_b = \text{background probability density}$

**Exercise 6:** Show that a binned multi-Poisson likelihood yields an un-binned likelihood of this form as the bin widths go to zero.

The likelihood function is arguably the most important quantity in a statistical analysis because it can be used to answer questions such as the following.
1. How do I estimate a parameter?
2. How do I quantify its accuracy?
3. How do I test a hypothesis?
4. How do I quantify the significance of a result?

Writing down the likelihood function requires:

1. identifying all that is known, e.g., the observations,
2. identifying all that is unknown, e.g., the parameters,
3. constructing a probability model for both.

Many analyses in particle physics do not use likelihood functions explicitly. However, since the data we use are stochastic, the failure to reflect deeply on their probabilistic nature and to model it explicitly leads to analyses that may not as good as they could be. Deconstructing carefully what is being done in an analysis is a habit that should be encouraged so that an accurate probabilistic model of the analysis can be constructed.

3 Lecture 2: the frequentist approach

In this lecture, we consider statistical inference from the frequentist viewpoint. In lecture 3, we consider the Bayesian approach. In our opinion, both are needed to make sense of statistical inference, though this is not the dominant opinion in particle physics.

The most important principle in the frequentist approach is that enunciated by the Polish statistician Jerzy Neyman in the 1930s, namely,

The frequentist principle

The goal of a frequentist analysis is to construct statements so that a fraction \( f \geq p \) of them are guaranteed to be true over an infinite ensemble of statements.

The fraction \( f \) is called the coverage probability, or coverage for short, and \( p \) is called the confidence level (C.L.). A procedure which satisfies the frequentist principle is said to cover. The confidence level as well as the coverage is a property of the ensemble of statements. Consequently, the confidence level may change if the ensemble changes. In a seminal paper published in 1937, Neyman [13] invented the concept of the confidence interval, a way to quantify uncertainty, that respects the frequentist principle. The confidence interval is such an important idea that it is worth working through the concept in detail.

3.1 Confidence intervals

Consider an experiment that observes \( D \) events with expected (that is, mean) signal \( s \) and no background. Neyman devised a way to make statements of the form

\[
s \in [l(D), u(D)],
\]

with the a priori guarantee that at least a fraction \( p \) of them will be true over an ensemble of statements of this kind. A procedure for constructing such intervals is called a Neyman construction. The frequentist principle must hold for any ensemble of experiments, not necessarily all making the same kind of observations and statements. For simplicity, however, we shall consider the experiments to be of the same kind and to be completely specified by a single unknown parameter \( s \). The Neyman construction is illustrated in Fig. 3.

The construction proceeds as follows. Choose a value of \( s \) and use some rule to find an interval in the space of observations (or, more generally, a region), for example, the interval defined by the two
Fig. 3: The Neyman construction. Plotted is the Cartesian product of the parameter space, with parameter \( s \), and the space of observations with potential observations \( D \). For a given value of \( s \), the observation space is partitioned into three disjoint intervals, such that the probability to observe a count \( D \) within the interval demarcated by the two vertical lines is \( f \geq p \), where \( p = \text{C.L.} \) is the desired confidence level. The inequality is needed because, for discrete data, it may not be possible to find an interval with \( f = p \) exactly.

vertical lines in the center of the figure, such that the probability to obtain a count in this interval is \( f \geq p \), where \( p \) is the desired confidence level. Then move to another value of \( s \) and repeat the procedure. The procedure is repeated for a sufficiently dense set of points in the parameter space over a sufficiently large range in \( s \). When this is done, as illustrated in Fig. 3, the intervals of probability content \( f \) will form a band in the Cartesian product of the parameter space and the observation space. The upper edge of this band defines the curve \( u(D) \), while the lower edge defines the curve \( l(D) \). These curves are the outcome of the Neyman construction.

For a given value of \( s \), the interval with probability content \( f \) in the space of observations is not unique since different rules for choosing the interval will, in general, yield different intervals. Neyman suggested choosing the interval so that the probability to obtain an observation to the right or left of the interval are the same (for a given value of \( s \)), which yields the so-called central intervals. One virtue of these intervals is that their boundaries can be more efficiently calculated by solving the equations,

\[
P(x \leq D | u) = \alpha_L, \quad P(x \geq D | l) = \alpha_R,
\]

a mathematical fact that becomes clear if we stare at Fig. 3 long enough.

Another rule was suggested by Feldman and Cousins [14]. For our example, the Feldman-Cousins rule requires that the potential observations \( \{D\} \) be ordered in descending order, \( D(1), D(2), \ldots \), of the likelihood ratio \( p(D|s)/p(D|\hat{s}) \), where \( \hat{s} \) is the maximum likelihood estimator (see Sec. 3.2) of the parameter \( s \). Once ordered, we compute the running sum \( f = \sum_j p(D(j)|s) \) until \( f \) equals or just exceeds the desired confidence level \( p \). This rules does not guarantee that the potential observations \( D \) are contiguous, but this does not matter because we simply take the minimum element of the set \( \{D(j)\} \) to be the lower bound of the interval and its maximum element to be the upper bound.

Another simple rule is the mode-centered rule: order \( D \) in descending order of \( p(D|s) \) and proceed as with the Feldman-Cousins rule. In principle, absent criteria for choosing a rule, there is nothing to prevent the use of ordering rules randomly chosen for different values of \( s \)!

Figure 4 compares the widths of the intervals \([l(D), u(D)]\) for three different ordering rules, central, Feldman-Cousins, and mode-centered as a function of the count \( D \). It is instructive to compare these widths with those provided
by the well-known root(N) interval, \( l(D) = D - \sqrt{D} \) and \( u(D) = D + \sqrt{D} \). Of the three sets of intervals, the ones suggested by Neyman are the widest, the Feldman-Cousins and mode-centered ones are of similar width, while the root(N) intervals are the shortest. So why are we going through all the trouble of the Neyman construction? We shall return to this question shortly.

Having completed the Neyman construction and found the curves \( u(D) \) and \( l(D) \) we can use the latter to make statements of the form \( s \in [l(D), u(D)] \): for a given observation \( D \), we simply read off the interval \( [l(D), u(D)] \) from the curves. For example, suppose in Fig. 3 that the true value of \( s \) is represented by the horizontal line that intersects the curves \( u(D) \) and \( l(D) \) and which therefore defines the interval demarcated by the two vertical lines. If the observation \( D \) happens to fall in the interval to the left of the left vertical line, or to the right of the right vertical line, then the interval \( [l(D), u(D)] \) will not bracket \( s \). However, if \( D \) falls between the two vertical lines, the interval \( [l(D), u(D)] \) will bracket \( s \). Moreover, by virtue of the Neyman construction, a fraction \( f \) of the intervals \( [l(D), u(D)] \) will bracket the value of \( s \) whatever its value happens to be, which brings us back to the question about the root(N) intervals. Figure 5 shows the coverage probability over the parameter space of \( s \). As expected, the three rules, Neyman’s, that of Feldman-Cousins, and the mode-centered, satisfy the condition coverage probability \( \geq \) confidence level over all values of \( s \) that are possible \textit{a priori}; that is, the intervals cover. However, the root(N) intervals do not and indeed fail badly for \( s < 2 \).

![Fig. 4: Interval widths as a function of count \( D \) for four sets of intervals.](image)

![Fig. 5: Interval widths as a function of count \( D \) for four sets of intervals.](image)

However, the coverage probability of the root(N) intervals bounces around the (68%) confidence level for values of \( s > 2 \). Therefore, if we knew for sure that \( s > 2 \), it would seem that using the root(N) intervals may not be that bad after all.

So what, after all this, does the statement \( s \in [l(D), u(D)] \) at 100\( p \)% C.L. mean in this approach, given that \( p \) is a property of the ensemble to which this statement belongs? In means this: \( s \in [l(D), u(D)] \) is a member of an ensemble of statements a fraction \( f \geq p \) of which are true. In principle, in order to verify this we need just count how many statements of the form \( s \in [l(D), u(D)] \) are true and divide by the total number of statements. Unfortunately, this requires that we know which statements are true.

But if we knew that we would not need a theory of statistical inference!

Neyman required a procedure to cover whatever the value of \textit{all} the parameters, be they known or unknown, of the probability models that describe the data generation mechanisms. This is a very tall order, which cannot be met in general. In practice, we resort to approximations, the most widely used of which is the profile likelihood to which we now turn.
3.2 The profile likelihood

As noted in Section 2.3, likelihood functions can be used to estimate the parameters on which they depend. The method of choice to do so, in a frequentist analysis, is called **maximum likelihood**, a method first used by Karl Frederick Gauss and developed into a formidable statistical tool in the 1930s by Sir Ronald A. Fisher [15], perhaps the most influential statistician of the twentieth century. The DØ top quark discovery example illustrates the method.

**Example: Top Quark Discovery Revisited**

We start by listing

the knowns

\[ D = N, B \]

where

\[ N = 17 \] observed events
\[ B = 3.8 \] estimated background events with uncertainty \( \delta B = 0.6 \)

and the unknowns

\[ b \] mean background count
\[ s \] mean signal count.

Next, we construct a probability model for the data \( D = N, B \). Since this is a counting experiment, we shall assume that \( p(x|s, b) \) includes a Poisson distribution with mean count \( s + b \). In the absence of details about how the background \( B \) was arrived at, the standard assumption is that data of the form \( y \pm \delta y \) can be modeled with a Gaussian (or normal) density. However, we shall do something slightly better.

Suppose that the observed count in the control region is \( Q \) and the mean count is \( bk \), where \( k \) (ideally) is the known scale factor between the control and signal regions. But, since we are given \( B \) and \( \delta B \) rather than \( Q \) and \( k \), we need to relate the two pairs of numbers. The simplest model is \( B = Q/k \) and \( \delta B = \sqrt{Q/k} \) from which we can infer an effective count \( Q \) using \( Q = (B/\delta B)^2 \). Since the scale factor \( k \) is not given, we shall use the obvious estimate \( k \sim Q/B = (B/\delta B)^2 \). With these assumptions, our likelihood function is

\[
p(D|s, b) = \text{Poisson}(N, s + b) \text{Poisson}(Q, bk),
\]

where

\[
Q = (B/\delta B)^2 = 41.11,
\]
\[
k = B/\delta B^2 = 10.56.
\]

The first term in Eq. (15) is the likelihood for the count \( N = 17 \), while the second term is the likelihood for \( B = 3.8 \), or equivalently the count \( Q \). The fact that \( Q \) is not an integer causes no difficulty; we merely continue the Poisson distribution to non-integer \( Q \) using \( (bk)^Q \exp(-bk)/\Gamma(Q + 1) \).

The maximum likelihood estimators for \( s \) and \( b \) are found by maximizing Eq. (15), that is, by solving the equations

\[
\frac{\partial \ln p(D|s, b)}{\partial s} = 0 \quad \text{leading to } \hat{s} = N - B,
\]
\[
\frac{\partial \ln p(D|s, b)}{\partial b} = 0 \quad \text{leading to } \hat{b} = B,
\]
as expected.

A more complete model would account for the uncertainty in \( k \).
The maximum likelihood method is the most widely used method for estimating parameters because it generally leads to reasonable estimates. But the method has features, or encourages practices, which, somewhat uncharitably, we label the good, the bad, and the ugly!

- **The Good**
  - Maximum likelihood estimators are consistent: the RMS goes to zero as more and more data are included in the likelihood. This is an extremely important property, which basically says it makes sense to take more data because we shall get more accurate results. One would not knowingly use an inconsistent estimator!
  - If an unbiased estimator for a parameter exists the maximum likelihood method will find it.
  - Given the MLE for \( s \), the MLE for any function \( y = g(s) \) of \( s \) is, very conveniently, just \( \hat{y} = g(\hat{s}) \). This is a very nice practical feature which makes it possible to maximize the likelihood using the most convenient parameterization of it and then transform back to the parameter of interest at the end.

- **The Bad (according to some!)**
  - In general, MLEs are biased.

**Exercise 7:** Show this

Hint: Taylor expand \( y = g(\hat{s} + h) \) about the MLE \( \hat{s} \), then consider its ensemble average.

- **The Ugly (according to some!)**
  - The fact that most MLEs are biased encourages the routine application of bias correction, which can waste data and, sometimes, yield absurdities.

Here is an example of the seriously ugly.

**Example**

For a discrete probability distribution \( p(k) \), the **moment generating function** is the ensemble average

\[
G(x) = \langle e^{xk} \rangle = \sum_k e^{xk} p(k).
\]

For the binomial, with parameters \( p \) and \( n \), this is

\[
G(x) = (e^x p + 1 - p)^n, \quad \text{Exercise 8a: Show this}
\]

which is useful for calculating **moments**

\[
\mu_r = \frac{d^r G}{dx^r} \bigg|_{x=0} = \sum_k k^r p(k),
\]

e.g., \( \mu_2 = (np)^2 + np - np^2 \) for the binomial distribution. Given that \( k \) events out \( n \) pass a set of cuts, the MLE of the event selection efficiency is the obvious estimate \( \hat{p} = k/n \). The equally obvious estimate of \( p^2 \) is \( (k/n)^2 \). But,

\[
\langle (k/n)^2 \rangle = p^2 + V/n, \quad \text{Exercise 8b: Show this}
\]
so \((k/n)^2\) is a biased estimate of \(p^2\) with positive bias \(V/n\). The unbiased estimate of \(p^2\) is

\[
k(k-1)/(n(n-1)),
\]

[Exercise 8c: Show this]

which, for a single success, i.e., \(k = 1\), yields the sensible estimate \(\hat{p} = 1/n\), but the less than useful \(\hat{p}^2 = 0!\)

In order to infer a value for the parameter of interest, for example, the signal \(s\) in our 2-parameter likelihood function in Eq. (15), the likelihood must be reduced to one involving the parameter of interest only, here \(s\), by getting rid of all the nuisance parameters, here the background parameter \(b\). A nuisance parameter is any parameter that is not of current interest. In a strict frequentist calculation, this reduction to the parameter of interest must be done in such a way as to respect the frequentist principle: coverage probability \(\geq\) confidence level. In general, this is very difficult to do exactly.

In practice, we replace all nuisance parameters by their conditional maximum likelihood estimates (CMLE). The CMLE is the maximum likelihood estimate conditional on a given value of the current parameter (or parameters) of interest. In the top discovery example, we construct an estimator of \(b\) as a function of \(s\), \(\hat{b}(s)\), and replace \(b\) in the likelihood \(p(D|s,b)\) by \(\hat{b}(s)\) to yield a function \(p_{PL}(D|s)\) called the profile likelihood.

Since the profile likelihood entails an approximation, namely, replacing unknown parameters by their conditional estimates, it is no longer the likelihood but rather an approximation to it. Consequently, the frequentist principle is not guaranteed to be satisfied exactly.

But, if certain conditions are met (Wilks’ theorem, 1938), roughly that the MLEs do not occur on the boundary of the parameter space and the likelihood becomes ever more Gaussian as the data become more numerous — that is, in the so-called asymptotic limit, then if the true density of \(x\) is \(p(x|s,b)\) the random number

\[
t(x,s) = -2 \ln \lambda(x,s),
\]

where

\[
\lambda(x,s) = \frac{p_{PL}(x|s)}{p_{PL}(x|\hat{s})},
\]

has a probability density that converges to a \(\chi^2\) density with one degree of freedom. More generally, if the numerator of \(\lambda\) contains \(m\) free parameters the asymptotic density of \(t\) is a \(\chi^2\) density with \(m\) degrees of freedom. Therefore, we may take \(t(D,s)\) to be a \(\chi^2\) variate, at least approximately, and solve \(t(D,s) = n^2\) for \(s\) to get approximate \(n\)-standard deviation confidence intervals. In particular, if we solve \(t(D,s) = 1\), we obtain approximate 68% intervals. This calculation is what Minuit, and now TMinuit, has done countless times since the 1970s! Wilks’ theorem provides the main justification for using the profile likelihood. We again use the top discovery example to illustrate the procedure.

**Example: Top Quark Discovery Revisited Again**

The conditional MLE of \(b\) is found to be

\[
\hat{b}(s) = \frac{g + \sqrt{g^2 + 4(1+k)Qs}}{2(1+k)},
\]

where

\[
g = N + Q - (1+k)s.
\]
The likelihood $p(D|s,b)$ is shown in Fig. 6(a) together with the graph of $\hat{b}(s)$. The mode (i.e. the peak) occurs at $s = \hat{s} = N - B$. By solving

$$-2 \ln \frac{p_{PL}(17|s)}{p_{PL}(17|17 - 3.8)} = 1$$

for $s$ we get two solutions $s = 9.4$ and $s = 17.7$. Therefore, we can make the statement $s \in [9.4, 17.7]$ at approximately 68% C.L. Figure 6(b) shows a plot of $-\ln \lambda(17, s)$ created using the RooFit [16] and RooStats [17] packages.

**Exercise 9:** Verify this interval using the RooFit/RooStats package

Intervals constructed this way are not guaranteed to satisfy the frequentist principle. In practice, however, their coverage is very good for the typical probability models used in particle physics, even for modest amounts of data.

### 3.3 Hypothesis tests

It is hardly possible in experimental particle physics to avoid testing hypotheses, testing that invariably leads to decisions. For example, electron identification entails hypothesis testing; given data $D$ we ask: is this particle an isolated electron or is it not an isolated electron? In the discovery of the Higgs boson, we had to test whether, given the data available in early summer 2012, the Standard Model without a Higgs boson, a somewhat ill-founded background-only model, or the Standard Model the new boson in July 2012, the background + signal model, was the preferred hypothesis. We decided that the latter model was preferred and announced the discovery of a new boson. Given the ubiquity of hypothesis testing, it is important to have a grasp of the methods that have been invented to implement it.

One method was due to Fisher [15], another was invented by Neyman, and a third (Bayesian) method was proposed by Sir Harold Jeffreys [18], all around the same time. We first describe the method of Fisher, then follow with a description of the method of Neyman. For concreteness, we consider the problem of deciding between a background-only model and a background + signal model.
3.3.1 Fisher’s approach

In Fisher’s approach, we construct a null hypothesis, often denoted by $H_0$, and reject it should some measure be judged small enough to cast doubt on the validity of this hypothesis. In our example, the null hypothesis is the background-only model, for example, the SM without a Higgs boson. The measure is called a p-value and is defined by

$$p\text{-value}(x_0) = P(x > x_0 | H_0),$$

(19)

where $x$ is a statistic designed so that large values indicate departure from the null hypothesis. This is illustrated in Fig. 7, which shows the location of the observed value $x_0$ of $x$. The p-value is the probability that $x$ could have been higher than the $x$ actually observed. It is argued that a small p-value implies that either the null hypothesis is false or something rare has occurred. If the p-value is extremely small, say $\sim 3 \times 10^{-7}$, then of the two possibilities the most common response is to presume the null to be false. If we apply this method to the DØ top quark discovery data, and neglect the uncertainty in the null hypothesis, we find

$$p\text{-value} = \sum_{N=17}^{\infty} \text{Poisson}(N, 3.8) = 5.7 \times 10^{-7}.$$ 

We usually report a more intuitive number by converting the p-value to the scale defined by

$$Z = \sqrt{2} \text{erf}^{-1}(1 - 2p\text{-value}).$$

(20)

This is the number of Gaussian standard deviations away from the mean. A p-value of $5.7 \times 10^{-7}$ corresponds to a $Z$ of 4.9$\sigma$. The $Z$-value can be calculated using the Root function

$$Z = -\text{TMath::NormQuantile(p-value)}.$$ 

3.3.2 Neyman’s approach

In Neyman’s approach two hypotheses are considered, the null hypothesis $H_0$ and an alternative hypothesis $H_1$. This is illustrated in Fig. 8. In our example, the null is the same as before but the alternative hypothesis is the SM with a Higgs boson. Again, one generally chooses $x$ so that large values would cast doubt on the validity of $H_0$. However, the Neyman test is specifically designed to respect the frequentist principle, which is done as follows. A fixed probability $\alpha$ is chosen called the significance (or size) of the test, which for a specific class of experiments corresponds to some threshold $x_\alpha$ defined by

$$\alpha = P(x > x_\alpha | H_0).$$

(21)

Should the observed value $x_0 > x_\alpha$, or equivalently, p-value($x_0$) < $\alpha$, the hypothesis $H_0$ is rejected in favor of the alternative. In particle physics, in addition to applying the Neyman hypothesis test, we also report the p-value. This is sensible because there is a more information in the p-value than merely reporting the fact that a null hypothesis was rejected at a significance level of $\alpha$.

---

\[^{1}\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{x} \exp(-t^2) \, dt \text{ is the error function.}\]
The Neyman method satisfies the frequentist principle by construction. Since the significance of the test is fixed, $\alpha$ is the relative frequency with which true null hypotheses would be rejected and is called the Type I error rate.

However, since we have specified an alternative hypothesis there is more that can be said. Figure 8 shows that we can also calculate

$$\beta = P(x \leq x_\alpha | H_1),$$

(22)

which is the relative frequency with which we would reject the hypotheses of the form of $H_1$ if they are true. These mistakes are called Type II errors. The quantity $1 - \beta$ is called the power of the test and is the relative frequency with which we would accept $H_1$ if true. Obviously, for a given $\alpha$ we want to maximize the power. Indeed, this is the basis of the Neyman-Pearson lemma (see for example Ref. [2]), which asserts that given two simple hypotheses — that is, hypotheses in which all parameters have well-defined values — the optimal statistic $t$ to use in the hypothesis test is the likelihood ratio $t = p(x|H_1)/p(x|H_0)$.

Maximizing the power seems sensible. Consider Fig. 9. The significance of the test in this figure is the same as that in Fig. 8, so the Type I error rate is identical. However, the Type II error rate is much greater in Fig. 9 than in Fig. 8, that is, the power of the test is considerably weaker in the former. In that case, there may be no compelling reason to reject the null since the alternative is not that much better. This insight was one source of Neyman’s disagreement with Fisher. Neyman objected to the possibility that one might reject a null hypothesis regardless of whether it made sense to do so. He insisted that the task is always one of deciding between competing hypotheses. Fisher’s counter argument was that an alternative hypothesis may not be available, but we may nonetheless wish to know whether the only hypothesis that is available is worth keeping. As we shall see, the Bayesian approach also requires an alternative, in agreement with Neyman, but in a way that neither he nor Fisher agreed with!

We have assumed that the hypotheses $H_0$ and $H_1$ are simple, that is, fully specified. Unfortunately, most of the hypotheses that arise in realistic particle physics analyses are not of this kind. In the Higgs boson discovery analyses by ATLAS and CMS the probability models depend on many nuisance parameters for which only estimates are available. Consequently, neither the background-only nor the background + signal hypotheses are fully specified. Such hypotheses are called compound hypotheses. In order to illustrate how hypothesis testing proceeds in this case, we again turn again to the top discovery example.

**Example**

As we saw in Sec. 3.2, the standard way to handle nuisance parameters in the frequentist approach is to replace them by their conditional MLEs and thereby reduce the likelihood function to the profile likelihood. In the top discovery example, we obtain a function $p_{PL}(D|s)$.
that depends on the single parameter, $s$. We now treat this function as if it were a likelihood and appeal to both the Neyman-Pearson lemma, which suggests the use of likelihood ratios, and Wilks’ theorem to motivate the use of the function $t(x, s)$ given in Eq. (17) to distinguish between two hypotheses: the hypothesis $H_1$ in which $s = \hat{s} = N - B$ and the hypothesis $H_0$ in which $s \neq \hat{s}$, for example, the background-only hypothesis $s = 0$. In the context of testing, $t(x, s)$ is called a test statistic, which, unlike a statistic as we have defined it (see Sec. 2.1), usually depends on at least one unknown parameter.

In principle, the next step is the computationally arduous task of simulating the distribution of the statistic $t(x, s)$. The task is arduous because a priori the probability density $p(t|s, b)$ can depend on all the parameters that exist in the original likelihood. If this is the case, then after all this effort we seem to have achieved a pyrrhic victory! But, this is where Wilks’ theorem saves the day, at least approximately. We can avoid the burden of simulating $t(x, s)$ because the latter is approximately a $\chi^2$ variate.

Using $N = 17$ and $s = 0$, we find $\sqrt{t_0} = \sqrt{t(17, 0)} = 4.6$. According to the results shown in Fig. (6)(a), $N = 17$ may can be considered “a lot of data”; therefore, we may use $t_0$ to implement a hypothesis test by comparing $t_0$ with a fixed value $t_\alpha$ corresponding to the significance level $\alpha$ of the test.

4 Lecture 3: the Bayesian approach

In this lecture, we introduce the Bayesian approach to inference, again using the top quark discovery data from DØ to illustrate the ideas.

The Bayesian approach is merely applied probability theory (see Section 2.2). A method is Bayesian if

– it is based on the degree of belief interpretation of probability and
– it uses Bayes theorem

$$p(\theta, \omega|D) = \frac{p(D|\theta, \omega) \pi(\theta, \omega)}{p(D)},$$

where

$D =$ observed data,

$\theta =$ parameters of interest,

$\omega =$ nuisance parameters,

$p(\theta, \omega|D) =$ posterior density,

$\pi(\theta, \omega) =$ prior density (or prior for short).

for all inferences. The result of a Bayesian inference is the posterior density $p(\theta, \omega|D)$ from which, if desired, various summaries can be extracted. The parameters can be discrete or continuous and nuisance parameters are eliminated by marginalization,

$$p(\theta|D) = \int p(\theta, \omega|D) d\omega,$$

$$\propto \int p(D|\theta, \omega) \pi(\theta, \omega) d\omega.$$ 

The function $\pi(\theta, \omega)$, called the prior, encodes whatever information we have about the parameters $\theta$ and $\omega$ independently of the data $D$. A key feature of the Bayesian approach is recursion: the use of the posterior density $p(\theta, \omega|D)$ or one, or more, of its marginals such as $p(\theta|D)$ as the prior in a subsequent analysis.
These simple rules yield an extremely powerful and general inference model, a model that was used, for example, in the discovery of single top quark production at the Tevatron [19, 20].

4.1 Model selection

Conceptually, hypothesis testing in the Bayesian approach (also called model selection) proceeds exactly the same way as any other Bayesian calculation: we compute the posterior density,

\[ p(\theta, \omega, H | D) = \frac{p(D | \theta, \omega, H) \pi(\theta, \omega, H)}{p(D)} , \]  

(25)

and marginalize it with respect to all parameters except the ones that label the hypotheses or models, \( H \),

\[ p(H | D) = \int p(\theta, \omega, H | D) d\theta d\omega. \]  

(26)

Equation (26) is the probability of hypothesis \( H \) given the observed data \( D \). In principle, the parameters \( \omega \) could also depend on \( H \). For example, suppose that \( H \) labels different parton distribution function (PDF) models, say CT10, MSTW, and NNPDF, then \( \omega \) would indeed depend on the PDF model and should be written as \( \omega_H \).

It is usually more convenient to arrive at the probability \( p(H | D) \) in stages.

1. Factorize the prior in the most convenient form,

\[ \pi(\theta, \omega_H, H) = \pi(\theta, \omega_H | H) \pi(H) , \]  

(27)

or

\[ = \pi(\theta | \omega_H, H) \pi(\omega_H | H) \pi(H) , \]  

(28)

Often, we can assume that the parameters of interest \( \theta \) are independent, \textit{a priori}, of both the nuisance parameters \( \omega_H \) and the model label \( H \), in which case we can write, \( \pi(\theta, \omega_H, H) = \pi(\theta) \pi(\omega_H | H) \pi(H) \).

2. Then, for each hypothesis, \( H \), compute the function

\[ p(D | H) = \int p(D | \theta, \omega_H, H) \pi(\theta, \omega | H) d\theta d\omega. \]  

(29)

3. Then, compute the probability of each hypothesis,

\[ p(H | D) = \frac{p(D | H) \pi(H)}{\sum_H p(D | H) \pi(H)} . \]  

(30)

Clearly, in order to compute \( p(H | D) \) it is necessary to specify the priors \( \pi(\theta, \omega | H) \) and \( \pi(H) \). With some effort, it is possible to arrive at an acceptable form for \( \pi(\theta, \omega | H) \), however, it is highly unlikely that consensus could ever be reached on the discrete prior \( \pi(H) \). At best, one may be able to adopt a convention. For example, if by convention two hypotheses \( H_0 \) and \( H_1 \) are to be regarded as equally likely, \textit{a priori}, then it would make sense to assign \( \pi(H_0) = \pi(H_1) = 0.5 \).

One way to circumvent the specification of the prior \( \pi(H) \) is to compare the probabilities,

\[ \frac{p(H_1 | D)}{p(H_0 | D)} = \left[ \frac{p(D | H_1)}{p(D | H_0)} \right] \frac{\pi(H_1)}{\pi(H_0)} . \]  

(31)

and use only the term in brackets, called the global \textbf{Bayes factor}, \( B_{10} \), as a way to compare hypotheses. The Bayes factor specifies by how much the relative probabilities of two hypotheses changes as a result of
incorporating new data, $D$. The word global indicates that we have marginalized over all the parameters of the two models. The local Bayes factor, $B_{10}(\theta)$ is defined by

$$B_{10}(\theta) = \frac{p(D|\theta, H_1)}{p(D|H_0)},$$

(32)

where,

$$p(D|\theta, H_1) \equiv \int p(D|\theta, \omega_{H_1}, H_1) \pi(\omega_{H_1}|H_1) d\omega_{H_1},$$

(33)

are the marginal or integrated likelihoods in which we have assumed the a priori independence of $\theta$ and $\omega_{H_1}$. We have further assumed that the marginal likelihood $H_0$ is independent of $\theta$, which is a very common situation. For example, $\theta$ could be the expected signal count $s$, while $\omega_{H_1} = \omega$ could be the expected background $b$. In this case, the hypothesis $H_0$ is a special case of $H_1$, namely, it is the same as $H_1$ with $s = 0$. An hypothesis that is a special case of another is said to be nested in the more general hypothesis. The example, discussed below, will make this clearer.

There is a notational subtlety that may be missed: because of the way we have defined $p(D|\theta, H)$, we need to multiply $p(D|\theta, H)$ by the prior $\pi(\theta)$ and then integrate with respect to $\theta$ in order to calculate $p(D|H)$.

### 4.1.1 Priors

Constructing a prior for nuisance parameters is generally neither controversial (for most parameters) nor problematic. The Achilles heal of the Bayesian approach is the need to specify the prior $\pi(\theta)$, for the parameters of interest, at the start of the inference chain when we know almost nothing about them. Careless specification of this prior can yield results that are unreliable or even nonsensical. The mandatory requirement is that the posterior density be proper, that is integrate to unity. Ideally, the same should hold for priors. A very extensive literature exists on the topic of prior specification when the available information is extremely limited. However, a discussion of this topic is beyond the scope of these lectures.

For model selection, we need to proceed with caution because Bayes factor are sensitive to the choice of priors and therefore less robust than posterior densities. Suppose that the prior $\pi(\theta) = C f(\theta)$, where $C$ is a normalization constant. The global Bayes factor for the two hypotheses $H_1$ and $H_0$ can be written as

$$B_{10} = C \int \frac{p(D|\theta, H_1) f(\theta) d\theta}{p(D|H_0)}.$$

(34)

Therefore, if the constant $C$ is ill defined, typically because $\int f(\theta) d\theta = \infty$, the Bayes factor will likewise be ill defined. For this reason, it is generally recommended that an improper prior not be used for parameters $\theta$ that occur only in one hypothesis, here $H_1$. However, for parameters that are common to all hypotheses, it is permissible to use improper priors because the constants cancel in the Bayes factor.

The discussion so far has been somewhat abstract. The next section therefore works through an example of a possible Bayesian analysis of the DØ top discovery data.

### 4.2 The top quark discovery: a Bayesian analysis

In this section, we shall perform the following calculations as a way to illustrate a typical Bayesian analysis,

1. compute the posterior density $p(s|D)$,
2. compute a 68% credible interval $[l(D), u(D)]$, and
3. compute the global Bayes factor $B_{10} = p(D|H_1)/p(D|H_0)$. 

212
**Probability model**

The first step in any serious statistical analysis is to think deeply about what has been done in the physics analysis and construct a probability model. The full probability model is the joint probability,

\[ p(x, s, b|I), \]

which, as is true of all probability models, is conditional on the information and assumptions, \( I \), that define the abstract space \( \Omega \) (see Sec. 2.2). In these lectures, we have omitted the conditioning data \( I \), but it should not be forgotten that it is always present and may differ from one probability model to another.

The full probability model \( p(x, s, b) \) can be factorized in several mathematically valid ways. However, we find it convenient to factorize the model in the following way,

\[ p(x, s, b) = p(x|s, b) \pi(s, b), \quad (35) \]

where we have introduced the symbol \( \pi \) in order to highlight the distinction we choose to make between this part of the model and the remainder. We shall compute the likelihood from \( p(x|s, b) \) and view \( \pi(s, b) \) as the prior for \( s \) and \( b \). We assume \( p(x|s, b) \) to be

\[ p(x|s, b) = \text{Poisson}(x, s + b). \quad (36) \]

The interpretation of \( p(x|s, b) \) is clear: it is the probability to observe \( x \) events given that the mean event count is \( s + b \). What does \( \pi(s, b) \) represent? This prior encodes what we know, or assume, about the mean background and signal independently of the potential observations \( x \). The prior \( \pi(s, b) \) can be factored in two ways,

\[ \pi(s, b) = \pi(s|b) \pi(b), \quad = \pi(b|s) \pi(s). \quad (37) \]

The factorizations remind us that the parameters \( s \) and \( b \) may not be independent. However, we shall assume that they are, at least at this stage of the analysis, in which case it is permissible to write,

\[ \pi(s, b) = \pi(s) \pi(b). \quad (38) \]

What do we know about the background? We know the count \( Q \) in the control region and we have an estimate of the control region to signal region scale factor \( k \). Since \( Q \) is a count, a reasonable model for the likelihood is

\[ p(Q|k, b) = \text{Poisson}(Q, kb), \quad (39) \]

from which, together with a prior \( \pi(k, b) \), we can compute the posterior density

\[ p(b|Q, k) = p(Q|k, b) \pi(k, b)/p(Q, k). \quad (40) \]

As usual, we factorize the prior, \( \pi(k, b) = \pi(k|b)\pi_0(b) \), where we have introduced the subscript 0 to distinguish \( \pi_0(b) \) from the background prior associated with Eq. (36). But, now we need to construct \( \pi(k|b) \) and \( \pi_0(b) \) using whatever information we have at hand.

Clearly, \( b \geq 0 \). But, that miserable tidbit is all we know apart from the background likelihood, Eq. (39)! Today, after a century of argument and discussion, the consensus amongst statisticians is that there is no unique way to represent such vague information. However, well founded ways to construct such priors are available, see for example Ref. [21] and references therein, but for simplicity we take the prior \( \pi_0(b) = 1 \), that is, the flat prior. If the uncertainty in \( k \) can be neglected, the (proper!) prior for \( k \) is \( \pi(k|b) = \delta(k - B/\delta B^2) \), which amounts to replacing \( k \) in Eq. (40) by \( B/\delta B^2 \). This yields,

\[ p(b|Q, k) = \text{Gamma}(kb, 1, Q + 1) = \frac{e^{-kb(kb)Q}}{\Gamma(Q + 1)}, \quad (41) \]

for the posterior density of \( b \), which can serve as the prior \( \pi(b) \) associated with Eq. (36).
By construction, \( p(x, s, b) \) is identical in form to the likelihood in Eq. (15); we have simply availed ourselves of the freedom to factorize \( p(x, s, b) \) as we wish and therefore to reinterpret the factors. This freedom is useful because it makes it possible to keep the likelihood simple while relegating the complexity to the prior. This may not seem, at first, to be terribly helpful; after all, we arrived at the same mathematical form as Eq. (15). However, the complexity can be substantially mitigated by sampling from the prior so that the model is represented by the relatively simple likelihood and an ensemble of points that collectively represent the prior. The likelihood, as we have conceptualized the problem, is given by

\[
p(D|s, b) = \frac{e^{-(s+b)}(s+b)^D}{D!},
\]

where \( D = 17 \) events.

The final ingredient is the prior \( \pi(s) \). At this stage, all we know is that \( s \geq 0 \) and, again, there is no unique way to specify \( \pi(s) \), though as noted there are well founded methods to construct it. We shall assume either the improper prior \( \pi(s) = 1 \) or the proper prior \( \pi(s) = \delta(s - 14) \).

**Marginal likelihood**

We have done the hard part: building the full probability model. Hereafter, the rest of the Bayesian analysis is mere computation.

It is convenient to eliminate the nuisance parameter \( b \),

\[
p(D|s, H_1) = \int_0^\infty p(D|s, b) \pi(b) d(b),
\]

\[
= \frac{1}{Q} (1-x)^2 \sum_{r=0}^N \text{Beta}(x, r+1, Q) \text{Poisson}(N-r|s),
\]

where \( x = 1/(1+k) \),

**Exercise 10:** Show this

and thereby arrive at the marginal likelihood \( p(D|s, H_1) \).

**Posterior density**

Given the marginal likelihood \( p(D|s, H_1) \) and a prior \( \pi(s) \) we can compute the posterior density,

\[
p(s|D, H_1) = p(D|s, H_1) \pi(s)/p(D|H_1),
\]

where,

\[
p(D|H_1) = \int_0^\infty p(D|s, H_1) \pi(s) ds.
\]

Fig. 10: Posterior density computed for DØ top quark discovery data. The shaded area is the 68% central credible interval.
Assuming a flat prior for the signal, $\pi(s) = 1$, we find

$$p(s|D, H_1) = \frac{\sum_{r=0}^{N} \text{Beta}(x, r + 1, Q) \text{Poisson}(N - r|s)}{\sum_{r=0}^{N} \text{Beta}(x, r + 1, Q)}, \quad (45)$$

Exercise 11: Derive an expression for $p(s|D, H_1)$ assuming $\pi(s) = \text{Gamma}(qs, 1, M + 1)$ where $q$ and $M$ are constants from which we can compute the central credible interval $[9.9, 18.4]$ for $s$ at 68% C.L., which is shown in Fig. 10. The statement $s \in [9.9, 18.4]$ at 68% C.L. means there is a 68% probability that $s$ lies in $[9.9, 18.4]$. Unlike the frequentist statement, this statement is about this particular interval and the 68% is a degree of belief, not a relative frequency. That being said, the best Bayesian methods tend to produce credible intervals that also approximate confidence intervals.

4.2.1 Bayes factor

As noted, the number $p(D|H_1)$ can be used to perform a hypothesis test. But, as argued above, we need to use a proper prior for the signal, that is, a prior that integrates to one. The simplest such prior is a $\delta$-function, e.g., $\pi(s) = \delta(s - 14)$. Using this prior, we find

$$p(D|H_1) = p(D|14, H_1) = 9.28 \times 10^{-2}.$$  

Since the background-only hypothesis $H_0$ is nested in $H_1$, and defined by $s = 0$, the number $p(D|H_0)$ is given by $p(D|0, H_1)$, which yields

$$p(D|H_0) = p(D|0, H_1) = 3.86 \times 10^{-6}.$$  

We conclude that the hypothesis $s = 14$ is favored over $s = 0$ by a Bayes factor of 24,000. In order to avoid large numbers, the Bayes factor can be mapped into a (signed) measure akin to the frequentist “$n$-sigma” [22],

$$Z = \text{sign}(\ln B_{10}) \sqrt{2|\ln B_{10}|}, \quad (46)$$

which gives $Z = 4.5$. Negative values of $Z$ correspond to hypotheses that are excluded.

Summary

We have given an overview of the main ideas of statistical inference in a form directly applicable to statistical analysis in particle physics. Two widely used approaches were covered, frequentist and Bayesian. Statistics is not physics. While Nature is the ultimate arbiter of which physics ideas are “correct”, the ultimate arbiter of statistical ideas is intellectual taste. Therefore, we hope you take to heart the following advice.

“Have the courage to use your own understanding”

Immanuel Kant

Acknowledgement

I thank Nick Ellis, Martijn Mulders, Helene Haller, and their counterparts from JINR, for organizing and hosting an excellent school, and the students for their energetic participation. These lectures were supported in part by US Department of Energy grant DE-FG02-97ER41022.
References


Cosmology and Gravitation: the grand scheme for High-Energy Physics

P. Binétruy
Université Paris Diderot, Paris, France

Abstract
These lectures describe how the Standard Model of cosmology (ΛCDM) has developed, based on observational facts but also on ideas formed in the context of the theory of fundamental interactions, both gravitational and non-gravitational, the latter being described by the Standard Model of high energy physics. It focuses on the latest developments, in particular the precise knowledge of the early Universe provided by the observation of the Cosmic Microwave Background and the discovery of the present acceleration of the expansion of the Universe. While insisting on the successes of the Standard Model of cosmology, we will stress that it rests on three pillars which involve many open questions: the theory of inflation, the nature of dark matter and of dark energy. We will devote one chapter to each of these issues, describing in particular how this impacts our views on the theory of fundamental interactions. More technical parts are given in italics. They may be skipped altogether.

1 A not so brief history of modern cosmology

Cosmology has been an enquiry of the human kind probably since the dawn of humanity. Modern cosmology was born in the early XXth century with the bold move of Einstein and contemporaries to apply the equations of general relativity, the theory of gravity, to the whole Universe. This has led to many successes and/or surprises, the most notable of which being presumably the discovery of extra-galactic objects which recede from our own Galaxy, i.e. the discovery of the expansion of the Universe [1, 2]. This led to the development of the Big Bang theory, with the early Universe being a hot and dense medium (a prediction confirmed by the discovery of the cosmic microwave background by Penzias and Wilson in 1965 [3]), and thus a laboratory for studying elementary particles. A picture thus emerged in the 1970s, not only based on the theory of gravity, but also on non-gravitational interactions described by the Standard Model of high energy physics, which was being finalized at the same time (its experimental confirmation would take another 40 years and has culminated in the discovery of the Higgs particle in 2012).

A first success of the particle physics approach to cosmology has been the understanding of the abundancy of light elements in the 80s. This was the first quantitative success of cosmology. Meanwhile, the development of gauge symmetries and the understanding of the rôle of spontaneous symmetry breaking in fundamental interactions led the community to focus its attention on phase transitions in the early Universe, in particular associated with the quark-gluon transition, the breaking of the electroweak symmetry or even of the grand unified symmetry. It is in this context that, in the early 80s, the theory of inflation was proposed [4, 5] to solve some of the mysteries of the standard Big Bang theory.

The theory of inflation included a model for the genesis of density fluctuations responsible for the formation of large scale structures, such as galaxies or clusters of galaxies: the quantum fluctuations during the exponential (de Sitter) expansion. But this implied the presence of fluctuations in the otherwise homogeneous and isotropic Cosmic Microwave Background. Such fluctuations were observed by the COBE satellite, at the level of one part in 100 000. Generic models of inflation predicted also in a very elegant manner that space (not spacetime!) is flat, any spatial curvature being erased by the exponential expansion. According to Einstein’s equations, this implied that the average energy density in the Universe had the critical value $\rho_c \sim 10^{-26}$ kg/m$^3$. 


This was a prediction not supported by observation. It was known since the 1930s that there was a significant amount of non-luminous –or dark– matter in the Universe: in 1933, Fritz Zwicky, by studying the velocity distribution of galaxies in the Coma cluster, had identified that there was 400 times more mass than expected from their luminosity. This had been confirmed by studying subsequently the rotation curves of many other galaxies. But the total of luminous and dark matter could not account for more than 30% of the critical energy density (other components like radiation are subdominant at present times). Models of open inflation were even constructed to reconcile inflation with observation.

The clue came in 1999 [6, 7] when it was observed that the expansion of the Universe is presently accelerating. Since matter or radiation tend to decelerate the expansion, one has to resort to a new form of energy, named dark energy, to understand this acceleration. Was this the component which would provide the missing 70% to account for a total energy density \( \rho_c \), and thus a spatially flat Universe?

The answer came from a more precise study of the fluctuations in the CMB through the space mission WMAP (and more recently Planck): they conclude indeed that these fluctuations are consistent with spatial flatness.

The latest cosmology results from the Planck mission, released this year, have confirmed the predictions of the simplest models of inflation, a rather remarkable feat since they are associated with dynamics active in the first fractions of seconds after the big bang, and they allow to fully understand the imprints observed 350,000 years after the big bang.

We thus have at our disposal a Standard Model of cosmology which is sometimes compared with the Standard Model of high energy physics: in both cases, no major experimental/observational data seems to be in conflict with the Model. There is however one big difference. The Standard Model of cosmology rests on three “pillars” –inflation, dark matter, dark energy– which are very poorly known: we have at present no microscopic theory of inflation, and we ignore the exact nature of dark matter or dark energy.

For example, there are convincing arguments that dark matter is made of weakly interacting massive particles of a new type, and a large experimental programme has been set up to identify them. Their discovery would be of utmost importance because this would be the first sign of physics beyond the Standard Model of particle physics. But it remains a possibility –though not a favored one– to explain the observed facts through a modification of gravity at different scales (from galaxies to clusters and cosmological scales). Finally, axion dark matter would be a minimal extension of the Standard Model, accounting for dark matter.

The discovery of the Higgs has provided us with the first example of a fundamental scalar field (at least fundamental at the scale where we observe it). This a welcome for cosmology since microscopic models of acceleration of the expansion of the Universe –whether inflation or dark energy– make heavy use of such fields. They have the double advantage of being non-vanishing without breaking the symmetries of spacetime (like Lorentz symmetry) and of having the potential of providing an unclustered background. They also appear naturally in the context of extensions of the Standard Model, like supersymmetry or extra dimensions.

Scalar fields thus provide valuable toy models of inflation or dark energy. However, such toys models are difficult to implement into realistic high energy physics models because of the constraints existing on physics beyond the Standard Models of particle physics and cosmology.

In what follows, we will review the successes of the Standard Model of cosmology, focussing on the most recent results. And we will consider the three pillars of this Model –inflation, dark matter, dark energy– and identify the most pressing questions concerning these three concepts.

In what follows some more technical discussions are given in italics. They can be skipped altogether. Background material is also provided in Appendices, most of it being intended for the reader who wants to reproduce some of the more advanced results.

We start in this first Chapter with an introduction, partly historical, to the main concepts of cos-
1.1 Gravity governs the evolution of the Universe

The evolution of the universe at large is governed by gravity, and thus described by Einstein’s equations. We recall that general relativity is based on the assumption made by Einstein that observations made in an accelerating reference are indistinguishable from those made in a gravitational field (as illustrated on a simple example in Fig. 1). This has several consequences, the most notable of which being that the curves followed by light (null geodesics) are not straight lines (see Fig. 2). Einstein’s equations are highly non-linear second order differential equations for the metric $g_{\mu\nu}$.

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 8\pi G_N T_{\mu\nu} + \lambda g_{\mu\nu}.$$  

(3)

where $R_{\mu\nu}$ is the Ricci tensor, $R$ the associated curvature scalar (see Appendix B) and $T_{\mu\nu}$ the energy-momentum tensor; finally $\lambda$ is the cosmological constant which has the dimension of an inverse length squared.

---

The metric signature we adopt throughout is Einstein’s choice: $(+,-,-,-)$.

These equations can be obtained from the Einstein-Hilbert action:

$$S = \frac{1}{8\pi G_N} \int d^4x \sqrt{-g} \left[ -\frac{1}{2} R - \lambda \right] + S_m(\psi, g_{\mu\nu}),$$

(2)

where the generic fields $\psi$ contribute to the energy-momentum: $T_{\mu\nu} = (2/\sqrt{-g})(\delta S_m/\delta g^{\mu\nu})$
Thus Einstein’s equations relate the geometry of spacetime (the left-hand side of (3)) with its matter field content (the right-hand side).

Einstein equations are field equations. Of which field? This is better understood in the weak gravitational field limit, that is in the limit of an almost flat spacetime. In this case, the metric is approximated by

$$g_{\mu \nu} \sim \eta_{\mu \nu} + h_{\mu \nu}(x),$$ (4)

where $\eta_{\mu \nu}$ is the Minkowski metric and $h_{\mu \nu}(x)$ is interpreted as the spin-2 graviton field. In particular, we note that

$$g_{00} = 1 + 2\Phi,$$ (5)

where $\Phi$ is the Newtonian potential which satisfies the Poisson equation $\Delta \Phi = 4\pi G_N \rho$.

Einstein used his equations not just to describe a given gravitational system like a planet or a star but the evolution of the whole universe. In his days, this was a bold move: it should be remembered how little of the universe was known at the time these equations were written. “In 1917, the world was supposed to consist of our galaxy and presumably a void beyond. The Andromeda nebula had not yet been certified to lie beyond the Milky Way.”[Pais [8] p. 286] Indeed, it is in this context that Einstein introduced the cosmological constant in order to have a static solution (until it was observed by Hubble that the universe is expanding) for the universe.

More precisely [8], Einstein first noticed that a slight modification of the Poisson equation, namely

$$\Delta \Phi - \lambda \Phi = 4\pi G_N \rho,$$ (6)

allowed a solution with a constant density $\rho$ ($\Phi = -4\pi G_N \rho / \lambda$) and thus a static Newtonian universe. In the context of general relativity, he found a static solution of (3) under the condition that

$$\lambda = \frac{1}{r^2} = 4\pi G_N \rho,$$ (7)
where \( r \) is the spatial curvature (see Exercise 1-1). It was soon shown that this Einstein universe is unstable to small perturbations.

### Exercise 1-1

Consider the following metric

\[
g_{00} = 1, \quad g_{ij} = -\delta_{ij} + \frac{x_i x_j}{x^2 - r^2}, \quad x^2 = \sum_{i=1}^{3} x_i^2.
\]  

(a) Show that it is a solution of Einstein’s equations (3) in the case of non-relativistic matter with a constant energy density \( \rho \) satisfying the condition (7).

(b) Prove that, in the Newtonian limit, one recovers (6).

Hints: 

(a) \( \Gamma^i_{jk} = \frac{1}{2} g^{il} \left( \partial_j g_{lk} + \partial_k g_{lj} - \partial_l g_{jk} \right) \) which gives \( R_{ij} = -2g_{ij}/r^2 \).

(b) In the Newtonian limit, \( G_{00} \sim \Delta g_{00} \) with \( g_{00} \) given by (4).

### 1.2 An Expanding Universe

Other solutions to the Einstein equations were soon discovered. The first exact non-trivial one was found in late 1915 by Schwarzschild, who was then fighting in the German army, within a month of the publication of Einstein’s theory and presented on his behalf by Einstein at the Prussian Academy in the first days of 1916 [9], just before Schwarzschild death from a illness contracted at the front. It describes static isotropic regions of empty spacetime (\( \lambda = 0 \)), such as the ones encountered in the exterior of a static star of mass \( M \) and radius \( R \):

\[
ds^2 = \left( 1 - \frac{2G_N M}{r} \right) dt^2 - \left( 1 - \frac{2G_N M}{r} \right)^{-1} dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2.
\]  

In 1917, de Sitter proposed a time-dependent vacuum solution in the case where \( \lambda \neq 0 \) [10, 11]:

\[
ds^2 = dt^2 - e^{2Ht} dx^2, \quad H^2 \equiv \frac{\lambda^2}{3}.
\]

But, since there exists time-dependent solutions, why should the Universe be static? This led to the so-called “Great Debate” between Harlow Shapley and Heber D. Curtis in 1920\(^3\). H. Shapley was supporting the view that the Universe is composed of only one big Galaxy, the spiral nebulae being just nearby gas clouds (he was also arguing –rightfully– that our Sun is far from the center of this big Galaxy). H.D. Curtis, on the other hand considered, that the Universe is made of many galaxies like our own, and that some of these galaxies had already been identified, in the form of the spiral nebulae (and he was supporting the view that our Sun is close to the centre of our relatively small Galaxy).

In 1925, Edwin Hubble studies, with the 100 inch Hooker telescope of Mount Wilson, the Cepheids which are variable stars in the Andromeda nebula M31. He shows that the distance is even greater than the size proposed by Shapley for our Milky Way: M31 is a galaxy of its own, the Andromeda galaxy (at a distance of \( 3.10^6 \) light-years from us) [12].

In 1929, Hubble [2] discovers, by combining spectroscopic measurements with measures of distance, that galaxies at a distance \( d \) from us recede at a velocity \( v \) following the law:

\[v \sim H_0 d,\]

the constant of proportionality being henceforth called the Hubble constant (see Fig. 3)\(^4\). As a consequence of the Hubble law, the Universe is expanding!

---

\(^3\)see http://apod.nasa.gov/diamond_jubilee/debate20.html

\(^4\)Hubble’s result was actually anticipated by G. Lemaître who published the law (11) two years earlier [1].
The velocities \( v \) are measured by Hubble through the Doppler shift of spectroscopic lines of the light emitted by the galaxy: if \( \lambda_{\text{emit}} \) is the wavelength of some spectroscopic line of the light emitted by a galaxy (receding from us at velocity \( v \)) and \( \lambda_{\text{obs}} \) of the corresponding line in the light observed by us, then
\[
1 + z \equiv \frac{\lambda_{\text{obs}}}{\lambda_{\text{emit}}} = \frac{1 + v/c}{\sqrt{1 - v^2/c^2}} \sim 1 + v/c .
\]
Hence \( v \simzc \).

Let us note that stars within galaxies, such are our Sun, are not individually subject to the expansion: they have fallen into the gravitational potential of the galaxy and are thus not receding from one another (see Exercise 1-2). This is why one had first to identify extragalactic objects before discovering the expansion: only galaxies, or clusters of galaxies, recede from one another.

**Exercise 1-2**: The purpose of this exercise is to show that, in the case where a fluctuation of density appears (such as when galaxies form), the massive objects (stars) decouple from the general expansion to fall into the local gravitational potential [14, 15].

In a matter-dominated universe of uniform density \( \rho \), a perturbation appears in the form of a sphere with uniform excess density \( \Delta \rho \). Assuming that the gravitational field inside the sphere is described by the Robertson-Walker metric with positive curvature constant \( \Delta k > 0 \), the evolution is governed by the equations
\[
\frac{\dot{a}^2}{a^2} + \frac{\Delta k}{a^2} = \frac{1}{3}(\rho + \Delta \rho) , \tag{13}
\]
\[
a^3(\rho + \Delta \rho) = C , \tag{14}
\]
where \( C \) is a constant and we neglect the cosmological constant.
a) Show that the solution is given parametrically by
\[ a(t) = \frac{C}{6\Delta k} (1 - \cos \eta), \quad (15) \]
\[ t = t_0 + \frac{C}{6\Delta k^{3/2}} (\eta - \sin \eta). \quad (16) \]
Note that this does not assume that \( \Delta \rho \) is small.

b) Show that, as \( t \to t_0 \), one has \( a(t) = (C/3)^{1/3} \left[ 3(t - t_0)/2 \right]^{2/3} \sim (t - t_0)^{2/3}, \) as in the rest of the matter-dominated universe. Verify that, whereas \( \rho = (4/3)(t - t_0)^{2/3} \sim \Delta \rho = (9/5)\Delta k a^{-2} \sim (t - t_0)^{-4/3}. \)

c) How long does it take before the system starts to collapse to a bound system or to a singularity?

Hints: a) \( a \ddot{a} + a \Delta k = C/3 \)

b) Make an expansion in \( \eta \) to the subleading order (\( \eta \) is small when \( t \to t_0 \)).

c) From (15) the expansion stops and the collapse starts at \( \eta = \pi \) or \( t - t_0 = \pi C/(6\Delta k^{3/2}) \).

Exercise 1-3: Identify the redefinition of coordinates \( \bar{t} = \bar{t}(t, r), \bar{r} = \bar{r}(t, r) \) which allows to write the de Sitter metric (10)
\[ ds^2 = dt^2 - e^{2Ht}(dr^2 + r^2 d\Omega^2), \quad (17) \]
into the following form:
\[ ds^2 = \left(1 - \frac{\bar{r}^2}{R_H^2}\right) dt^2 - \left(1 - \frac{\bar{r}^2}{R_H^2}\right)^{-1} d\bar{r}^2 - \bar{r}^2 d\Omega^2, \quad R_H \equiv H^{-1}. \quad (18) \]
The first equation is known as the flat form of the de Sitter metric (see (19) below), the second one is the static form (compare with the Schwarzschild metric (9)).

Hints: \( e^{Ht} = e^{H\bar{t}} \sqrt{1 - H^2 \bar{r}^2} \) and \( r = \bar{r} e^{-Ht} \).

1.3 Friedmann-Lemaître-Robertson-Walker universe

As one gets to larger and larger distances, the Universe becomes more homogeneous and isotropic. Under the assumption that it reaches homogeneity and isotropy on scales of order 100 Mpc (1 pc = 3.262 light-year = 3.086 \times 10^{16} m) and larger, one may first try to find a homogeneous and isotropic metric as a solution of Einstein’s equations. The most general ansatz is, up to coordinate redefinitions, the Robertson-Walker metric:
\[ ds^2 = c^2 dt^2 - a^2(t) \gamma_{ij} dx^i dx^j, \quad (19) \]
\[ \gamma_{ij} dx^i dx^j = \frac{dr^2}{1 - kr^2} + r^2 \left( d\theta^2 + \sin^2 \theta d\phi^2 \right), \quad (20) \]
where \( a(t) \) is the cosmic scale factor, which is time-dependent in an expanding or contracting universe. Such a universe is called a Friedmann-Lemaître universe. The constant \( k \) which appears in the spatial metric \( \gamma_{ij} \) can take the values ±1 or 0: the value 0 corresponds to flat space, i.e. usual Minkowski spacetime; the value +1 to closed space (\( r^2 < 1 \)) and the value −1 to open space. Note that \( r \) is dimensionless whereas \( a \) has the dimension of a length. From now on, we set \( c = 1 \), except otherwise stated.

For the energy-momentum tensor that appears on the right-hand side of Einstein’s equations, we follow our assumption of homogeneity and isotropy and assimilate the content of the Universe to a perfect fluid:
\[ T_{\mu\nu} = -p g_{\mu\nu} + (p + \rho) U_\mu U_\nu, \quad (21) \]
where $U^\mu$ is the velocity 4-vector ($U^t = 1, U^i = 0$). It follows from (21) that $T_{tt} = \rho$ and $T_{ij} = a^2 p \gamma_{ij}$. The pressure $p$ and energy density $\rho$ usually satisfy the equation of state:

\[ p = w \rho \]  \hspace{1cm} (22)

The constant $w$, called the equation of state parameter, takes the value $w \sim 0$ for non-relativistic matter (negligible pressure) and $w = 1/3$ for relativistic matter (radiation). In all generality, the perfect fluid consists of several components with different values of $w$.

One obtains from the $(0, 0)$ and $(i, j)$ components of the Einstein equations (3) (see Exercise B-1 of Appendix B):

\[ 3 \left( \frac{\dot{a}^2}{a^2} + \frac{k}{a^2} \right) = 8\pi G_N \rho + \lambda, \]  \hspace{1cm} (23)

\[ \dot{a}^2 + 2a\ddot{a} + k = -8\pi G_N a^2 \rho + a^2 \lambda, \]  \hspace{1cm} (24)

where we use standard notations: $\dot{a}$ is the first time derivative of the cosmic scale factor, $\ddot{a}$ the second time derivative.

The first of the preceding equations can be written as the Friedmann equation, which gives an expression for the Hubble parameter $H \equiv \dot{a}/a$ measuring the rate of the expansion of the Universe:

\[ H^2 = \frac{\dot{a}^2}{a^2} = \frac{1}{3} \left( \lambda + 8\pi G_N \rho \right) - \frac{k}{a^2}. \]  \hspace{1cm} (25)

The cosmological constant appears as a constant contribution to the Hubble parameter. We note that, setting $k = 0$ and $\rho = 0$, one recovers de Sitter solution (10): $\dot{a}^2/a^2 = \lambda/3$ i.e. $a(t) \sim e^{Ht}$. For the time being, we will set $\lambda$ to zero and return to it in subsequent chapters.

Next, we note that, assuming $k = 0$, we have at present time:

\[ \rho = \frac{3H_0^2}{8\pi G_N} \equiv \rho_c \sim 10^{-26} \text{kg/m}^3, \]  \hspace{1cm} (26)

where $H_0$ is the Hubble constant, i.e. the present value of the Hubble parameter. This corresponds to approximately one galaxy per Mpc$^3$ or 5 protons per m$^3$. In fundamental units where $\hbar = c = 1$, this is of the order of $(10^{-3}\text{eV})^4$. We easily deduce from (25) that space is open (resp. closed) if at present time $\rho < \rho_c$ (resp. $\rho < \rho_c$). Hence the name critical density for $\rho_c$.

Friedmann equation (25) can be understood on very simple grounds: since the universe at large scale is homogeneous and isotropic, there is no specific location and motion in the universe should not allow to identify any such location. This implies that the most general motion has the form $\mathbf{v}(t) = H(t) \mathbf{x}$ where $\mathbf{x}$ and $\mathbf{v}$ denote the position and the velocity and $H(t)$ is an arbitrary function of time. Since $\mathbf{v} = \mathbf{\dot{x}}$, one obtains $\mathbf{x} = a(t) \mathbf{r}$, where $\mathbf{r}$ is a constant for a given body (called the comoving coordinate) and $a(t)$ is related to $H(t)$ through $H = \dot{a}/a$. Now, consider a particle of mass $m$ located at position $\mathbf{x}$: the sum of its kinetic and gravitational potential energy is constant. Denoting by $\rho$ the energy density of the (homogeneous) universe, we have

\[ \frac{1}{2} m \mathbf{v}^2 - \frac{4\pi}{3} G_N m \rho |\mathbf{x}|^2 = \text{cst}, \]  \hspace{1cm} (27)

Writing this constant $-k m r^2/2$, we obtain from (27)

\[ \frac{\dot{a}^2}{a^2} = \frac{8\pi}{3} G_N \rho - \frac{k}{a^2}, \]  \hspace{1cm} (28)

which is nothing but Friedmann equation (25) (with vanishing cosmological constant).
Friedmann equation should be supplemented by the conservation of the energy-momentum tensor which simply yields:
\[ \dot{\rho} = -3H(p + \rho) \]  
(29)
Hence a component with equation of state (22) has its energy density scaling as \( \rho \sim a(t)^{-3(1+w)} \). Thus non-relativistic matter (often referred to as matter) energy density scales as \( a^{-3} \). In other words, the energy density of matter evolves in such a way that \( \rho a^3 \) remains constant. Radiation scales as \( a^{-4} \) and a component with equation of state \( p = -\rho \ (w = -1) \) has constant energy density\(^5\).

We note for future use that, if a component with equation of state (22) dominates the energy density of the universe (as well as the curvature term \(-k/a^2\)), then (28) has a scaling solution
\[ a(t) \sim t^\nu, \quad \text{with} \quad \nu = \frac{2}{3(1+w)} \]  
(30)
For example, in a matter-dominated universe, \( a(t) \sim t^{2/3} \), and in a radiation-dominated universe, \( a(t) \sim t^{1/2} \).

Differentiating the Friedmann equation with respect to time, and using the energy-momentum conservation (29), one easily obtains
\[ \ddot{a} = -\frac{4\pi G_N}{3}a(3p + \rho) + a\frac{\lambda}{3} \]  
(31)
This allows to recover (24) from Friedmann equation and energy-momentum conservation.

### 1.4 Redshift

In an expanding or contracting universe, the Doppler frequency shift undergone by the light emitted from a distant source gives a direct information on the time dependence of the cosmic scale factor \( a(t) \). To obtain the explicit relation, we consider a photon propagating in a fixed direction (\( \theta \) and \( \phi \) fixed). Its equation of motion is given as in special relativity by setting \( ds^2 = 0 \) in (19):
\[ c^2 dt^2 = a^2(t) \frac{dr^2}{1 - kr^2} \]  
(32)
Thus, if a photon (an electromagnetic wave) leaves at time \( t \) a galaxy located at distance \( r \) from us, it will reach us at time \( t_0 \) such that
\[ \int_t^{t_0} \frac{cdt}{a(t)} = \int_0^r \frac{dr}{\sqrt{1 - kr^2}} \]  
(33)
The electromagnetic wave is emitted with the same amplitude at a time \( t + T \) where the period \( T \) is related to the wavelength of the emitted wave \( \lambda \) by the relation \( \lambda = cT \). It is thus received with the same amplitude at the time \( t_0 + T_0 \) given by
\[ \int_{t+T}^{t_0+T_0} \frac{cdt}{a(t)} = \int_0^r \frac{dr}{\sqrt{1 - kr^2}} \]  
(34)
the wavelength of the received wave being simply \( \lambda_0 = cT_0 \). Since \( T_0, T \ll t_0, t \), we obtain from comparing (33) and (34)
\[ \frac{cT_0}{a_0} = \frac{cT}{a(t)} \quad \text{i.e.} \quad \frac{\lambda_0}{\lambda} = \frac{a_0}{a(t)} \]  
(35)
where \( a_0 \) is the present value of the cosmic scale factor.

Defining the redshift parameter \( z \) as the fractional increase in wavelength \( z = (\lambda_0 - \lambda)/\lambda \), we have
\[ 1 + z = \frac{a_0}{a(t)} \]  
(36)
One may thus replace time by redshift since time decreases monotonically as redshift increases.

\(^5\)The latter case corresponds to a cosmological constant as can be seen from (23-24) where the cosmological constant can be replaced by a component with \( \rho_\Lambda = -p_\Lambda = \lambda/(8\pi G_N) \).
1.5 The universe today: energy budget

The Friedmann equation
\[ H^2 = \frac{\dot{a}^2}{a^2} = \frac{1}{3} \left( \lambda + \frac{8\pi G_N \rho}{a^2} \right) - \frac{k}{a^2}. \]  
(37)
allows to define the Hubble constant \( H_0 \), i.e. the present value of the Hubble parameter, which sets the scale of our Universe at present time. Because of the troubled history of the measurement of the Hubble constant, it has become customary to express it in units of 100 km.s\(^{-1}\).Mpc\(^{-1}\) which gives its order of magnitude. Present measurements give
\[ h_0 = \frac{H_0}{100 \text{ km.s}^{-1}\text{Mpc}^{-1}} = 0.7 \pm 0.1. \]
The corresponding length and time scales are:
\[ \ell_H = \frac{c}{H_0} = 3000 h_0^{-1} \text{ Mpc} = 9.25 \times 10^{25} h_0^{-1} \text{ m}, \]
(38)
\[ t_H = \frac{1}{H_0} = 3.1 \times 10^{17} h_0^{-1} \text{ s} = 9.8 h_0^{-1} \text{ Gyr}. \]
(39)

It has become customary to normalize the different forms of energy density in the present Universe in terms of the critical density \( \rho_c = 3H_0^2/(8\pi G_N) \) defined in (26). Separating the energy density \( \rho_{\Delta 0} \) presently stored in non-relativistic matter (baryons, neutrinos, dark matter,...) from the density \( \rho_{R0} \) presently stored in radiation (photons, relativistic neutrino if any), one defines:
\[ \Omega_M \equiv \frac{\rho_{\Delta 0}}{\rho_c}, \quad \Omega_R \equiv \frac{\rho_{R0}}{\rho_c}, \quad \Omega_\Lambda \equiv \frac{\lambda}{3H_0^2}, \quad \Omega_k \equiv -\frac{k}{a_0^2H_0^2}. \]
(40)
The last term comes from the spatial curvature and is not strictly speaking a contribution to the energy density. One may add other components: we will refrain to do so in this Chapter and defer this to the last one.

Then the Friedmann equation taken at time \( t_0 \) simply reads
\[ \Omega_M + \Omega_R + \Omega_\Lambda = 1 - \Omega_k. \]
(41)
Since matter dominates over radiation in the present Universe, we may neglect \( \Omega_R \) in the preceding equation. As we will see in the next Chapters, present observational data tend to favor the following set of values: \( \Omega_k \sim 0 \) (see Section 2.1) and \( \Omega_M \sim 0.3, \Omega_\Lambda \sim 0.7 \) (see Section 5.1 and Fig. 22). The matter density is not consistent with the observed density of luminous matter (\( \Omega_{\text{luminous}} \sim 0.04 \)) and thus calls for a nonluminous form of matter, dark matter which will be studied in Section 4.

Using the dependence of the different components with the scale factor \( a(t) = a_0/(1 + z) \), one may then rewrite the Friedmann equation at any time as:
\[ H^2(t) = H_0^2 \left[ \Omega_\Lambda + \Omega_M \left( \frac{a_0}{a(t)} \right)^3 + \Omega_R \left( \frac{a_0}{a(t)} \right)^4 + \Omega_k \left( \frac{a_0}{a(t)} \right)^2 \right], \]
(42)
or \[ H^2(z) = H_0^2 \left[ \Omega_M (1 + z)^3 + \Omega_R (1 + z)^4 + \Omega_k (1 + z)^2 + \Omega_\Lambda \right]. \]
(43)
where \( a_0 \) is the present value of the cosmic scale factor and all time dependences (or alternatively redshift dependence) have been written explicitly. We note that, even if \( \Omega_M \) is negligible in (41), this is not so in the early Universe because the radiation term increases faster than the matter term in (42) as one gets back in time (i.e. as \( a(t) \) decreases). If we add an extra component \( X \) with equation of state \( p_X = w_X \rho_X \), it contributes an extra term \( \Omega_X \left( \frac{a_0}{a(t)} \right)^{3(1+w_X)} \) where \( \Omega_X = \rho_X/\rho_c \).
An important information about the evolution of the universe at a given time is whether its expansion is accelerating or decelerating. The acceleration of our universe is usually measured by the deceleration parameter \( q \) which is defined as:

\[
q \equiv -\frac{\ddot{a}}{a^2}.
\]

Using (31) of Section 2 and separating again matter and radiation, we may write it at present time \( t_0 \) as:

\[
q_0 = -\frac{1}{H_0^2} \left( \frac{\ddot{a}}{a} \right)_{t=t_0} = \frac{1}{2} \Omega_M + \Omega_R - \Omega_\Lambda.
\]

Once again, the radiation term \( \Omega_R \) can be neglected in this relation. We see that in order to have an acceleration of the expansion \( (q_0 < 0) \), we need the cosmological constant to dominate over the other terms.

We can also write the deceleration parameter in (44) in terms of redshift as in (43)

\[
q = \frac{H_0^2}{2H(z)^2} \left[ \Omega_M (1+z)^3 + 2 \Omega_R (1+z)^4 - 2 \Omega_\Lambda \right].
\]

This shows that the universe starts accelerating at redshift values \( 1+z \sim (2\Omega_\Lambda/\Omega_M)^{1/3} \) (neglecting \( \Omega_R \)), that is typically redshifts of order 1.

The measurements of the Hubble constant and of the deceleration parameter at present time allow to obtain the behaviour of the cosmic scale factor in the last stages of the evolution of the universe:

\[
a(t) = a_0 \left[ 1 + \frac{t - t_0}{t_H} - \frac{q_0}{2} \left( \frac{t - t_0}{t_H} \right)^2 + \cdots \right].
\]

### 1.6 The early universe

Table 1 summarizes the history of the universe in the context of the big bang model with an inflationary epoch. We are referring to the different stages using time, redshift or temperature. The last two can be related using the conservation of entropy.

Indeed, one can show, using the second law of thermodynamics \( (TdS = dE + pdV) \) that the entropy per unit volume is simply the quantity

\[
s \equiv \frac{S}{V} = \frac{\rho + p}{T}.
\]

and that the entropy in a covolume \( sa^3 \) remains constant. The entropy density is dominated by relativistic particles and reads

\[
s = \frac{2}{3} g_s(T) a_B T^3,
\]

\[
g_s(T) = \sum_{\text{bosons}} g_i \left( \frac{T_i}{T} \right)^3 + \frac{7}{8} \sum_{\text{fermions}} g_i \left( \frac{T_i}{T} \right)^3,
\]

where the sum extends only to the species in thermal equilibrium and \( a_B \equiv \pi^2 k^4/(15c^3\hbar^3) = 7.56 \times 10^{-16} \text{ J} \cdot \text{m}^{-3} \cdot \text{K}^{-4} \) is the blackbody constant.

We deduce from the constancy of \( sa^3 \) that \( g_s \left( aT \right)^2 \) remains constant. Hence the temperature \( T \) of the universe behaves as \( a^{-1} \) whenever \( g_s \) remains constant. We conclude that as we go back in time (\( a \) decreases), temperature increases, as well as energy density. The early universe is hot and dense. It might even reach a stage where our equations no longer apply because it becomes infinitely hot and dense: this
is the initial singularity, sarcastically (but successfully) called big bang by Fred Hoyle, on a BBC radio show in 1949.

We note that some caution has to be paid whenever some species drop out of thermal equilibrium. Indeed, a given species drops out of equilibrium when its interaction rate $\Gamma$ drops below the expansion rate $H$. For example neutrinos decouple at temperatures below 1 MeV. Their temperature continues to decrease as $a^{-1}$ and thus remains equal to $T$. However when $kT$ drops below $2m_e$, electrons annihilate against positrons with no possibility of being regenerated and the entropy of the electron-positron pairs is transferred to the photons. Since $g_s|_{\gamma,e^\pm} = 2 + 4 \cdot 7/8 = 11/2$ and $g_s|_\gamma = 2$, the temperature of the photons becomes multiplied by a factor $(11/4)^{1/3}$. Since the neutrinos have already decoupled, they are not affected by this entropy release and their temperature remains untouched. Thus we have

$$\frac{T}{T_\nu} = \left(\frac{11}{4}\right)^{1/3} \approx 1.40 \quad .$$

We can then compute the value of $g_s$ for temperatures much smaller than $m_e$: $g_s = 2 + (7/8)6(4/11) = 3.91$. We deduce that, at present time ($T_0 = 2.725$ K), $s_0/k = 2890 \text{ cm}^{-3}$.

As we have seen in (30), it follows from the Friedmann equation that, if the Universe is dominated by a component of equation of state $p = w\rho$, then the cosmic scale factor $a(t)$ varies with time as $t^{2/[3(1+w)]}$. We start at time $t_0$ with the energy budget: $\Omega_M = 0.3, \Omega_\Lambda = 0.7, \Omega_k \sim 0$ (see next Chapters). Radiation consists of photons and relativistic neutrinos. Since generically

$$\rho_R = \frac{1}{2} g(T) a_H^4 T^4 \quad ,$$

where $g(T)$ is the effective number of degrees of freedom

$$g(T) = \sum_{\text{bosons}} g_i \left(\frac{T_i}{T}\right)^4 + \frac{7}{8} \sum_{\text{fermions}} g_i \left(\frac{T_i}{T}\right)^4 \quad .$$

(we have taken into account the possibility that the species $i$ may have a thermal distribution at a temperature $T_i$ different from the temperature $T$ of the photons), we have

$$\rho_R(t) = \rho_\gamma(t) \left[ 1 + \frac{7}{8} \left(\frac{4}{11}\right)^{4/3} N_{\nu, \text{rel}}^{\gamma}(t) \right] \quad ,$$

where $N_{\nu, \text{rel}}(t)$ is the number of relativistic neutrinos at time $t$. At present time $t_0$, we have $\Omega_\gamma = \rho_\gamma(t_0)/\rho_c = 2.48 \times 10^{-5} \text{ h}^{-2}$ and the mass limits on neutrinos imply $N_{\nu, \text{rel}}^{\gamma}(t_0) \leq 1$. In any case, $\Omega_R \ll \Omega_M$.

For redshifts larger than 1, the cosmological constant becomes subdominant and the universe is matter-dominated ($a(t) \sim t^{2/3}$). In the early phase of this matter-dominated epoch, the Universe is a ionized plasma with electrons and protons: it is opaque to photons. But, at a time $t_{\text{rec}}$, electrons recombine with the protons to form atoms of hydrogen and, because hydrogen is neutral, this induces the decoupling of matter and photon: from then on ($t_{\text{rec}} < t < t_0$), the universe is transparent. This is the important recombination stage. After decoupling the energy density $\rho_\gamma \sim T^4$ of the primordial photons is redshifted according to the law

$$\frac{T(t)}{T_0} = \frac{a_0}{a(t)} = 1 + z \quad .$$

After recombination, the intergalactic medium remains neutral during a period often called the dark ages, until the first stars ignite and the first quasars are formed. The ultraviolet photons produced by these sources progressively then re-ionize the universe. This period, called the re-ionization period, may be long since only small volumes around the first galaxies start to be ionized until these volumes coalesce to re-ionize the full intergalactic medium. But, in any case, the universe is then sufficiently dilute to prevent recoupling.
Table 1: The different stages of the cosmological evolution in the standard scenario, given in terms of time \( t \) since the big bang singularity, the energy \( kT \) of the background photons and the redshift \( z \). The double line following nucleosynthesis indicates the part of the evolution which has been tested through observation. The values \( (h_0 = 0.7, \Omega_{\text{m}} = 0.3, \Omega_{\Lambda} = 0.7) \) are adopted to compute explicit values.

<table>
<thead>
<tr>
<th>( t )</th>
<th>( kT ) (eV)</th>
<th>( z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_0 \sim 15 ) Gyr</td>
<td>( 2.35 \times 10^{-4} )</td>
<td>0</td>
</tr>
<tr>
<td>( \sim ) Gyr</td>
<td>( \sim 10^{-3} )</td>
<td>( \sim 10 )</td>
</tr>
<tr>
<td>( t_{\text{rec}} \sim 4 \times 10^5 ) yr</td>
<td>0.26</td>
<td>1100</td>
</tr>
<tr>
<td>( t_{\text{eq}} \sim 4 \times 10^4 ) yr</td>
<td>0.83</td>
<td>3500</td>
</tr>
<tr>
<td>3 min</td>
<td>( 6 \times 10^4 )</td>
<td>( 2 \times 10^8 )</td>
</tr>
<tr>
<td>1 s</td>
<td>( 10^6 )</td>
<td>( 3 \times 10^9 )</td>
</tr>
<tr>
<td>( 4 \times 10^{-6} ) s</td>
<td>( 4 \times 10^8 )</td>
<td>( 10^{12} )</td>
</tr>
<tr>
<td>( &lt; 4 \times 10^{-6} ) s</td>
<td>( &gt; 10^9 )</td>
<td></td>
</tr>
<tr>
<td>( t = 0 )</td>
<td>( \infty )</td>
<td></td>
</tr>
</tbody>
</table>

One observes presently this cosmic microwave background (CMB) as a radiation with a black-body spectrum at temperature \( T_0 = 2.725 \) K or energy \( kT_0 = 2.35 \times 10^{-4} \) eV.

Since the binding energy of the ground state of atomic hydrogen is \( E_b = 13.6 \) eV, one may expect that the energy \( kT_{\text{rec}} \) is of the same order. It is substantially smaller because of the smallness of the ratio of baryons to photons \( \eta = n_b/n_\gamma \sim 5 \times 10^{-10} \). Indeed, according to the Saha equation, the fraction \( x \) of ionized atoms is given by

\[
\frac{n_p n_e}{n_H n_\gamma} = \frac{x^2}{(1-x)^{\eta}} = \frac{4.05e^3}{\pi^2} \left( \frac{m_e}{2\pi kT} \right)^{3/2} e^{-E_b/kT} \tag{56}
\]
Hence, because $\eta \ll 1$, the ionized fraction $x$ becomes negligible only for energies much smaller than $E_b$. A careful treatment gives $kT_{rec} \sim 0.26$ eV.

As we proceed back in time, radiation energy density increases more rapidly (as $a(t)^{-4}$) than matter ($a(t)^{-3}$) (as $a(t)$ decreases). At time $t_{eq}$, there is equality. This corresponds to

$$\frac{1}{1 + z_{eq}} = \frac{a(t_{eq})}{a_0} = \frac{1.68 \Omega}{\Omega_M} = \frac{4.17 \times 10^{-5}}{\Omega_M h_0^2},$$

where we have assumed 3 relativistic neutrinos at this time.

As we go further back in time, we presumably reach a period where matter overcame antimatter. It is indeed a great puzzle of our Universe to observe so little antimatter, when our microscopic theories treat matter and antimatter on equal footing. More quantitatively, one has to explain the following very small number:

$$\eta \equiv \frac{n_B}{n_{\gamma}} = \frac{n_b - n_{\bar{b}}}{n_{\gamma}} \sim 6 \times 10^{-10},$$

where $n_B = n_b - n_{\bar{b}}$ (resp. $n_{\gamma}$) is the baryon (resp. photon) number density, based on baryon $b$ and antibaryon $\bar{b}$ counts. The actual number comes from the latest Planck data [16].

Sakharov [17] gave in 1967 the necessary ingredients to generate an asymmetry between matter and antimatter:

- a process that destroys baryon number,
- a violation of the symmetry between matter and antimatter (the so-called charge conjugation), as well as a violation of the time reversal symmetry,
- an absence of thermal equilibrium.

The Standard Model ensures the second set of conditions (CP violation which was discovered by Cronin and Fitch [18] is accounted for by the phase of the CKM matrix). The expanding early universe provides the third condition. It remains to find a process that destroys baryon number. Different roads were followed: non-perturbative processes (sphaleron) at the electroweak phase transition; proton decay in the context of grand unified theories; decay of heavy neutrinos which leads to lepton number violation, and consequently to baryon number violation (leptogenesis).

2 The days where cosmology became a quantitative science: cosmic microwave background

We have recalled briefly in Section 1.6 the history of the Universe (see Table 1). The very early universe is a ionized plasma, and thus is opaque to light. But we have seen that, soon after matter-radiation equality, electrons recombined with the protons to form neutral atoms of hydrogen, which induces the decoupling of matter and photon. From this epoch on, the universe becomes transparent to light. The primordial gas of photons produced at this epoch cools down as the universe expands (following 55) and forms nowadays the cosmic microwave background (CMB).

Bell Labs radio astronomers Arno Penzias and Robert Wilson were using a large horn antenna in 1964 and 1965 to map signals from the Milky Way, when they accidently discovered the CMB [3]. The discovery of this radiation was a major confirmation of the hot big bang model: its homogeneity and isotropy was a signature of its cosmological origin. However, the degree of homogeneity and isotropy of this radiation was difficult to reconcile with the history of the Universe as understood from the standard big bang theory: radiation coming from regions of the sky which were not supposed to be causally connected in the past had exactly the same properties.

It was for such reasons that the scenario of inflation was proposed [4,5]: an exponential expansion of the Universe right after the big bang. In the solution proposed by A. Guth [5], the set up was the
spontaneous breaking of the grand unified theory: the corresponding phase transition was providing the vacuum energy necessary to initiate such an exponential expansion. This scenario proved to be difficult to realize and it was followed by many variants: new inflation [19], chaotic inflation [20], ...

We have said that, for \( t < t_{rec} \), i.e. \( z > 1100 \), the Universe is a ionized plasma, opaque to electromagnetic radiation. This means that, when we observe the early Universe, we hit a “wall” at the corresponding redshift: the earlier Universe appears to our observation as a blackbody, and should thus radiate according to the predictions of Planck. This is certainly the largest blackbody that one could think of. This blackbody spectrum of the CMB was indeed observed [21] by the FIRAS instrument onboard the COsmic Background Explorer (COBE) satellite launched in 1989 by NASA (see Fig. 4).

More precisely, the CMB has the nearly perfect thermal spectrum of a black body at temperature \( T_\gamma = 2.725 \) K (corresponding to a number density \( n_\gamma = 411 \text{ cm}^{-3} \)):

\[
d\rho_\gamma = 2hf\frac{1}{e^{hf/kT_\gamma} - 1} \frac{4\pi f^2 df}{c^3}
\]

(the first factor accounts for the two polarizations) or

\[
\frac{d\rho_\gamma}{d\ln f} = 3.8 \times 10^{-15} \text{J/m}^3 \left( \frac{f}{f_\gamma} \right)^4 \frac{e^{-1}}{e^{f/f_\gamma} - 1},
\]

where \( f_\gamma = kT_\gamma/h = 5.7 \times 10^{10} \) Hz.

Another expectation for COBE was the presence of fluctuations in the CMB. Indeed, if inflation was to explain the puzzle of isotropy and homogeneity of the CMB through the whole sky, one expected that quantum fluctuations produced during the inflation phase would show up to some degree as tiny fluctuations of temperature in the CMB. Such fluctuations were discovered by the instrument DMR (Differential Microwave Radiometers) onboard COBE, at the level of one part to \( 10^5 \) [22] (see Fig. 5).

It is primarily homogeneous and isotropic but includes fluctuations at a level of \( 10^{-5} \), which are of much interest since they are.

Since the days of COBE, there has been an extensive study of the CMB fluctuations to identify the imprints of the recombination and earlier epochs. This uses the results of ground, balloon or space missions (WMAP in the US and most recently Planck in Europe). We will review this in some details in the next Section.

Fig. 4: Blackbody spectrum of the CMB as observed by the instrument FIRAS onboard the COBE satellite [21]
Fig. 5: Temperature fluctuations in the CMB observed by COBE [22] (the galaxy is removed). The galaxy as well as the dipole component associated to the Doppler effect due to the motion of the Earth have been removed. Blue areas are colder than average, red areas are warmer but no $7^\circ$ region varies from the mean by more than 200 $\mu$K ($\Delta T/T \sim 8.10^{-5}$).

2.1 CMB

Before discussing the spectrum of CMB fluctuations, we introduce the important notion of a particle horizon in cosmology.

Because of the speed of light, a photon which is emitted at the big bang ($t = 0$) will have travelled a finite distance at time $t$. The proper distance (C.1) measured at time $t$ is simply given by the integral:

\[ d_{ph}(t) = \frac{a(t)}{H_0} \int_{0}^{t} \frac{cdt'}{a(t')} \]

\[ = \frac{\ell H_0}{1 + z} \int_{z}^{\infty} \frac{dz}{[\Omega_M (1 + z)^3 + \Omega_R (1 + z)^4 + \Omega_k (1 + z)^2 + \Omega_\Lambda]^{1/2}}, \tag{61} \]

where, in the second line, we have used (C.2). This is the maximal distance that a photon (or any particle) could have travelled at time $t$ since the big bang. In other words, it is possible to receive signals at a time $t$ only from comoving particles within a sphere of radius $d_{ph}(t)$. This distance is known as the particle horizon at time $t$.

A quantity of relevance for our discussion of CMB fluctuations is the horizon at the time of the recombination i.e. $z_{rec} \sim 1100$. We note that the integral on the second line of (61) is dominated by the lowest values of $z$: $z \sim z_{rec}$ where the universe is still matter dominated. Hence

\[ d_{ph}(t_{rec}) \sim \frac{2\ell H_0}{\Omega_{M}^{1/2} z_{rec}^{3/2}} \sim 0.3 \text{ Mpc}. \tag{62} \]

One may introduce also the Hubble radius

\[ R_H(t) = H^{-1}(z), \tag{63} \]
which will play an important role in the following discussion. This scale characterizes the curvature of spacetime at the time \( t \) (see for example Exercise B.1.b). We note that the particle horizon is simply twice the Hubble radius at recombination, as can be checked from (43):

\[
R_H(t_{\text{rec}}) \sim \frac{\ell H_0}{\Omega_M^{1/2} z_{\text{rec}}^{3/2}}.
\]  

(64)

This radius is seen from an observer at present time under an angle

\[
\theta_H(t_{\text{rec}}) = \frac{R_H(t_{\text{rec}})}{d_A(t_{\text{rec}})},
\]

(65)

where the angular distance has been defined in (C.7). We can compute analytically this angular distance under the assumption that the universe is matter dominated (see Exercise C-1). Using (C.10), we have

\[
d_A(t_{\text{rec}}) = \frac{a_0 r}{1 + z_{\text{rec}}} \sim \frac{2\ell H_0}{\Omega_M z_{\text{rec}}^{1/2}}.
\]

(66)

Thus, since, in our approximation, the total energy density \( \Omega_T \) is given by \( \Omega_M \),

\[
\theta_H(t_{\text{rec}}) \sim \Omega_T^{1/2}/(2z_{\text{rec}}^{1/2}) \sim 0.015 \text{ rad} \Omega_T^{1/2} \sim 1^\circ \Omega_T^{1/2}.
\]  

(67)

We have written in the latter equation \( \Omega_T \) instead of \( \Omega_M \) because numerical computations show that, in case where \( \Omega_\Lambda \) is non-negligible, the angle depends on \( \Omega_M + \Omega_\Lambda = \Omega_T \).

We can now discuss the evolution of photon temperature fluctuations. For simplicity, we will assume a flat primordial spectrum of fluctuations: this leads to predictions in good agreement with experiment; moreover, as we will see in the next Section, it is naturally explained in the context of inflation scenarios.

Before decoupling, the photons are tightly coupled with the baryons through Thomson scattering. In a gravitational potential well, gravity tends to pull this baryon-photon fluid down the well whereas radiation pressure tends to push it out. Thus, the fluid undergoes a series of acoustic oscillations. These oscillations can obviously only proceed if they are compatible with causality i.e. if the corresponding wavelength is smaller than the horizon scale or the Hubble radius: \( \lambda = 2\pi a(t)/k < R_H(t) \) or

\[
k > 2\pi \frac{a(t)}{R_H(t)} \sim t^{-1/3}.
\]

(68)

Starting with a flat primordial spectrum, we see that the first oscillation peak corresponds to \( \lambda \sim R_H(t_{\text{rec}}) \), followed by other compression peaks at \( R_H(t_{\text{rec}})/n \) (see Fig. 6). They correspond to an angular scale on the sky:

\[
\theta_n \sim \frac{R_H(t_{\text{rec}})}{d_A(t_{\text{rec}})} \frac{1}{n} = \frac{\theta_H(t_{\text{rec}})}{n}.
\]

(69)

Since photons decouple at \( t_{\text{rec}} \), we observe the same spectrum presently (up to a redshift in the photon temperature)\(^7\).

Experiments usually measure the temperature difference of photons received by two antennas separated by an angle \( \theta \), averaged over a large fraction of the sky. Defining the correlation function

\[
C(\theta) = \left\langle \frac{\Delta T}{T_0} (\mathbf{n}_1) \frac{\Delta T}{T_0} (\mathbf{n}_2) \right\rangle
\]

(70)

\(^7\)A more careful analysis indicates the presence of Doppler effects besides the gravitational effects that we have taken into account here. Such Doppler effects turn out to be non-leading for odd values of \( n \).
averaged over all \( n_1 \) and \( n_2 \) satisfying the condition \( n_1 \cdot n_2 = \cos \theta \), we have indeed

\[
\left\langle \left( \frac{T(n_1) - T(n_2)}{T_0} \right)^2 \right\rangle = 2 \left( C(0) - C(\theta) \right).
\]

(71)

We may decompose \( C(\theta) \) over Legendre polynomials:

\[
C(\theta) = \frac{1}{4\pi} \sum_{l}^{\infty} (2l + 1)C_l P_l(\cos \theta).
\]

(72)

The monopole \((l = 0)\), related to the overall temperature \( T_0 \), and the dipole \((l = 1)\), due to the Solar system peculiar velocity, bring no information on the primordial fluctuations. A given coefficient \( C_l \) characterizes the contribution of the multipole component \( l \) to the correlation function. If \( \theta \ll 1 \), the main contribution to \( C_l \) corresponds to an angular scale\(^{8}\) \( \theta \sim \pi/l \sim 200^\circ/l \). The previous discussion (see (67) and (69)) implies that we expect the first acoustic peak at a value \( l \sim 200\Omega_T^{-1/2} \).

The power spectrum obtained by the Planck experiment is shown in Fig. 7. One finds the first acoustic peak at \( l \sim 200 \), which constrains the \( \Lambda \)CDM model used to perform the fit to \( \Omega_T = \Omega_M + \Omega_\Lambda \sim 1 \). Many other constraints may be inferred from a detailed study of the power spectrum [16, 23].

### 2.2 Baryon acoustic oscillations

We noted in the previous section that, before decoupling, baryons and photons were tightly coupled and the baryon-photon fluid underwent a series of acoustic oscillations, which have left imprints in the CMB: the characteristic distance scale is the sound horizon, which is the comoving distance that sound waves could travel from the Big Bang until recombination at \( z = z_* \):

\[
r_s = \int_{0}^{t_*} \frac{c_s(t)}{1 + z} dt = \int_{z_*}^{\infty} \frac{c_s(z)}{H(z)} dz,
\]

(73)

\(^{8}\)The \( C_l \) are related to the coefficients \( a_{lm} \) in the expansion of \( \Delta T/T \) in terms of the spherical harmonics \( Y_{lm} \): \( C_l = \langle |a_{lm}|^2 \rangle_m \). The relation between the value of \( l \) and the angle comes from the observation that \( Y_{lm} \) has \((l - m)\) zeros for \(-1 < \cos \theta < 1\) and \( \text{Re}(Y_{lm}) \) \( m \) zeros for \( 0 < \phi < 2\pi \).
where \( c_s \) is the sound velocity. This distance has been recently measured with precision by the Planck collaboration to be \( r_s = 144.96 \pm 0.66 \) Mpc [16].

We have until now followed the fate of photons after decoupling. Similarly, once baryons decouple from the radiation, their oscillations freeze in, which leads to specific imprints in the galaxy power spectrum, such as the characteristic scale \( r_s \). Indeed, remember that, until recombination, baryons and photons were tightly coupled (but not dark matter). Thus a given matter density perturbation may have travelled a distance \( r_s \) in the case of baryons under the influence of radiation pressure (to which the photons are sensitive), whereas it did not move in the case of dark matter. This will lead, once (dark and baryonic) matter has collapsed into galaxies, to a secondary peak a distance \( r_s \) away in the distribution of separations of pairs of galaxies. In 2005, Eisenstein and collaborators [24], using data from the Sloan Digital Sky Survey, have indeed identified such a baryon acoustic peak in the matter power spectrum (Fourier transform of the two-point correlation function) on scales of order \( 105 h^{-1} \sim 150 \) Mpc. Figure 41 shows how the \( \Lambda \)CDM model just described fares with respect to observations by comparison with a model with \( \Omega \Lambda = 0 \), a small value of the Hubble parameter and a small fraction of matter which does not cluster on small scale (relic neutrinos or quintessence).

The acoustic peak provides a standard ruler which can be used for measuring distances: measurements along the line of sight depend on \( H(z)r_s \) whereas measurements transverse to the line of sight depend on the angular diameter distance \( d_A(z)/r_s \). In fact, because the analysis rests on spherically-averaged two-point statistics, the distance scale determined is \( d_V \) defined as:

\[
\left( \frac{d_V(z)}{r_s} \right)^3 = \left( \frac{d_A(z)}{r_s} \right)^2 \frac{cz}{H(z)r_s}.
\]

Measuring the acoustic scale at \( z = 0 \) provides a standard distance ruler which allows to identify \( H_0 \),

**Fig. 7:** This figure compares the temperature spectrum for the best fit \( \Lambda \)CDM model, in red, with the temperature angular power spectrum observed by the Planck collaboration (in blue, averaged over bins of width \( \Delta \ell \sim 31 \) with \( 1 \sigma \) errors). The gray dots are the unbinned data. In the lower panel, the green lines show the \( \pm 1 \sigma \) errors on the individual power spectrum estimates. [16]
and then $\Omega_{\Lambda}$ (from $\Omega_{\Lambda} h^2$). Going to higher $z$, this allows to put constraints on the recent history of the Universe, and thus on the evolution of the dark matter component.

### 2.3 Inflation

The inflation scenario has been proposed to solve a certain number of problems faced by the cosmology of the early universe [5]. Among these one may cite:

- **the flatness problem**
  
  If the total energy density $\rho_T$ of the universe is presently close to the critical density, it should have been even more so in the primordial universe. Indeed, we can write (25) of Section 2 as

  \[
  \frac{\rho_T(t)}{\rho_c(t)} - 1 = \frac{k}{a^2},
  \]  

  where $\rho_c(t) = 3H^2(t)/(8\pi G_N)$ and the total energy density $\rho_T$ includes the vacuum energy. If we take for example the radiation-dominated era where $a(t) \sim t^{1/2}$, then (75) can be written as ($\dot{a} \sim t^{-1/2} \sim a^{-1}$)

  \[
  \frac{\rho_T(t)}{\rho_c(t)} - 1 = \left[ \frac{\rho_T(t_U)}{\rho_c(t_U)} - 1 \right] \left( \frac{a(t)}{a(t_U)} \right)^2 = \left[ \frac{\rho_T(t_U)}{\rho_c(t_U)} - 1 \right] \left( \frac{kT}{kT_U} \right)^2,
  \]

  where we have used the fact that $T(t) \propto a(t)^{-1}$ and we have taken as a reference point the epoch $t_U$ of the grand unification phase transition. This means that, if the total energy density is close to the critical density at matter-radiation equality (as can be inferred from the present value), it must be even more so at the time of the grand unification phase transition: by a factor $(1\,\text{eV}/10^{16}\,\text{GeV})^2 \sim 10^{-50}$! Obviously, the choice $k = 0$ in the spatial metric ensures $\rho_T = \rho_c$ but the previous estimate shows that this corresponds to initial conditions which are highly fine tuned.

- **the horizon problem**
We have stressed in the previous Sections the isotropy and homogeneity of the cosmic microwave background and identified its primordial origin. It remains that the horizon at recombination is seen on the present sky under an angle of $2^{\circ}$. This means that two points opposite on the sky were separated by about 100 horizons at the time of recombination, and thus not causally connected. It is then extremely difficult to understand why the cosmic microwave background should be isotropic and homogeneous over the whole sky.

- the monopole problem

Monopoles occur whenever a simple gauge group is broken to a group with a $U(1)$ factor. This is precisely what happens in grand unified theories. In this case their mass is of order $M_{\nu}/g^2$ where $g$ is the value of the coupling at grand unification. Because we are dealing with stable particles with a superheavy mass, there is a danger to overclose the universe, i.e. to have an energy density much larger than the critical density. We then need some mechanism to dilute the relic density of monopoles.

Inflation provides a remarkably simple solution to these problems: it consists in a period of the evolution of the universe where the expansion is exponential. Indeed, if the energy density of the universe is dominated by the vacuum energy $\rho_{\text{vac}}$ (or by some constant form of energy), then the Friedmann equation reads

$$H^2 = \frac{\dot{a}^2}{a^2} = \frac{\rho_{\text{vac}}}{3m_P^2}.$$  \hspace{1cm} (77)

where $m_P \equiv (8\pi G_N)^{-1/2}$ is the reduced Planck mass. If $\rho_{\text{vac}} > 0$, this is readily solved as

$$a(t) = H_{\text{vac}}^{-1} e^{H_{\text{vac}} t} \quad \text{with} \quad H_{\text{vac}} \equiv \sqrt{\frac{\rho_{\text{vac}}}{3m_P^2}}.$$  \hspace{1cm} (78)

Such a behaviour is in fact observed whenever the magnitude of the Hubble parameter changes slowly with time i.e. is such that $|\dot{H}| \ll H^2$.

As we have seen in (10), such a space was first proposed by de Sitter [10, 11] with very different motivations and is thus called de Sitter space.

Obviously a period of inflation will ease the horizon problem. Indeed, the particle horizon size during inflation reads, following (61)

$$d_{\text{ph}}(t)|_{\text{de Sitter}} = a(t) \int_{t_i}^{t} \frac{cdt'}{a(t')} = \frac{c}{H_{\text{vac}}} e^{H_{\text{vac}}(t-t_i)} \text{ for } H_{\text{vac}}(t-t_i) \gg 1.$$  \hspace{1cm} (79)

It follows that a period of inflation extending from $t_i$ to $t_f = t_i + \Delta t$ contributes to the particle horizon size a value $c e^{H_{\text{vac}} \Delta t}/H_{\text{vac}}$, which can be very large.$^9$

We note that de Sitter space also has a finite event horizon. This is the maximal distance that comoving particles can travel between the time $t$ where they are produced and $t = \infty$ (compare with (61):

$$d_{\text{eh}}(t) = a(t) \int_{t}^{\infty} \frac{cdt'}{a(t')}.$$  \hspace{1cm} (80)

In the case of de Sitter space, this is simply

$$d_{\text{eh}}(t)|_{\text{de Sitter}} = \frac{c}{H_{\text{vac}}} = R_H,$$  \hspace{1cm} (81)

i.e. it corresponds to the Hubble radius (constant for de Sitter spacetime). This allows to make an analogy between de Sitter spacetime and a black hole: we will see in Section 3.2 that a Schwarzschild

---

$^9$We also note that, in a pure de Sitter space, the particle horizon diverges as we take $t_i \to -\infty$. This reflects the fact that, in a de Sitter space, all points were in causal contact.
black hole of mass $M$ has an event horizon at the Schwarzschild radius $R_S = 2G_N M$ (see also Exercise 1-3 of Section 1 for a comparison between the Schwarzschild and the de Sitter metric in its static form). Thus, just as black holes evaporate by emitting radiation at Hawking temperature $T_H = 1/(4\pi R_S)$ (see Eq. (136)), an observer in de Sitter spacetime feels a thermal bath at temperature $T_H = H/(2\pi)$.

We see that it is the event horizon that fixes here the cut-off scale of microphysics. Since it is equal here to Hubble radius, and since the Hubble radius is of the order of the particle horizon for matter or radiation-dominated universe, it has become customary to compare the comoving scale associated to physical processes with the Hubble radius (we already did so in our discussion of acoustic peaks in CMB spectrum; see Fig. 6, and Fig. 9 below).

A period of exponential expansion of the universe may also solve the monopole problem by diluting the concentration of monopoles by a very large factor. It also dilutes any kind of matter. Indeed, a sufficiently long period of inflation “empties” the universe. However matter and radiation may be produced at the end of inflation by converting the energy stored in the vacuum. This conversion is known as reheating (because the temperature of the matter present in the initial stage of inflation behaves as $a^{-1}(t) \propto e^{-H_{vac}t}$, it is very cold at the end of inflation; the new matter produced is hotter). If the reheating temperature is lower than the scale of grand unification, monopoles are not thermally produced and remain very scarce.

Finally, it is not surprising that the universe comes out very flat after a period of exponential inflation. Indeed, the spatial curvature term in the Friedmann equation is then damped by a factor $a^{-2} \propto e^{-2H_{vac}\Delta t}$. For example, a value $H_{vac}\Delta t \sim 60$ (one refers to it as 60 $e$-foldings) would easily account for the huge factor $10^{50}$ of adjustment that we found earlier.

Most inflation models rely on the dynamics of a scalar field in its potential. Inflation occurs whenever the scalar field evolves slowly enough in a region where the potential energy is large. The set up necessary to realize this situation has evolved with time: from the initial proposition of Guth [5] where the field was trapped in a local minimum to “new inflation” with a plateau in the scalar potential [19, 26], chaotic inflation [20] where the field is trapped at values much larger than the Planck scale and more recently hybrid inflation [27] with at least two scalar fields, one allowing an easy exit from the inflation period.

The equation of motion of a homogeneous scalar field $\phi(t)$ with potential $V(\phi)$ evolving in a Friedmann-Robertson-Walker universe is:

$$\ddot{\phi} + 3H\dot{\phi} = -V'(\phi),$$

(82)

where $V'(\phi) \equiv dV/d\phi$. The term $3H\dot{\phi}$ is a friction term due to the expansion. The corresponding energy density and pressure are:

$$\rho = \frac{1}{2} \dot{\phi}^2 + V(\phi),$$

(83)

$$p = \frac{1}{2} \dot{\phi}^2 - V(\phi).$$

(84)

We may note that the equation of conservation of energy $\dot{\rho} = -3H(\rho + p)$ takes here simply the form of the equation of motion (82). These equations should be complemented with the Friedmann equation (77).

When the field is slowly moving in its potential, the friction term dominates over the acceleration term in the equation of motion (82) which reads:

$$3H\dot{\phi} \simeq -V'(\phi).$$

(85)

---

10In an open or flat universe, the event horizon is infinite.
The curvature term may then be neglected in the Friedmann equation (77) which gives
\[ H^2 \simeq \frac{\rho}{3m_p^2} \simeq \frac{V}{3m_p^2}. \] (86)

Then the equation of conservation \( \dot{\rho} = -3H(p + \rho) = -3H\dot{\phi}^2 \) simply gives
\[ \dot{H} \simeq -\frac{\dot{\phi}^2}{2m_p^2}. \] (87)

It is easy to see that the condition \( |\dot{H}| \ll H^2 \) amounts to \( \dot{\phi}^2/2 \ll \rho/3 \sim V(\phi)/3 \), i.e. a kinetic energy for the scalar field much smaller than its potential energy. Using (85) and (86), the latter condition then reads
\[ \epsilon \equiv \frac{1}{2} \left( \frac{m_pV'}{V} \right)^2 \ll 1. \] (88)

The so-called slow roll regime is characterized by the two equations (85) and (86), as well as the condition (88). It is customary to introduce another small parameter:
\[ \eta \equiv \frac{m_p^2V''}{V} \ll 1, \] (89)

which is easily seen to be a consequence of the previous equations\(^{11,12}\).

An important quantity to be determined is the number of Hubble times elapsed during inflation. From some arbitrary time \( t \) to the time \( t_e \) marking the end of inflation (i.e. of the slow roll regime), this number is given by
\[ N(t) = \int_t^{t_e} H(t) dt. \] (92)

It gives the number of e-foldings undergone by the scale factor \( a(t) \) during this period (see (78). Since \( dN = -Hdt = -Hd\phi/\dot{\phi} \), one obtains from (85) and (86)
\[ N(\phi) = \int_{\phi_e}^{\phi} \frac{1}{m_p^2} \frac{V}{V'} d\phi. \] (93)

During the inflationary phase, the scalar fluctuations of the metric may be written in a conformal Newtonian coordinate system as:
\[ ds^2 = a^2 \left[ (1 + 2\Phi)d\eta^2 - (1 - 2\Phi)\delta_{ij}dx^i dx^j \right], \] (94)
where \( \eta \) is conformal time (\( ad\eta = dt = da/\dot{a} \)). We may write the correlation function in Fourier space \( P_S(k) \) by
\[ \langle \Phi_k \Phi_{k'}^* \rangle = 2\pi^2 k^{-3} P_S(k) \delta^3 (k - k') \] (95)

The origin of fluctuations is found in the quantum fluctuations of the scalar field during the de Sitter phase. Indeed, if we follow a given comoving scale \( a(t)/k \) with time (see Fig. 9), we have seen in Section 2.1 that, some time during the matter-dominated phase, it enters the Hubble radius. Since \( a(t) \)

\(^{11}\)Differentiating (85), one obtains
\[ \eta = \epsilon - \ddot{\phi}/(H\dot{\phi}). \] (90)

\(^{12}\)Note that one finds also in the literature the slow roll coefficients defined from the Hubble parameter [28]
\[ \epsilon_H \equiv 2m_p^2 \left( \frac{H'(\phi)}{H(\phi)} \right)^2 = \epsilon, \ \eta_H \equiv 2m_p^2 \frac{H''(\phi)}{H(\phi)} = \eta - \epsilon. \] (91)
Fig. 9: Evolution of a physical comoving fluctuation scale with respect to the Hubble radius during the inflation phase \( R_H(t) = H_{\text{vac}}^{-1} \), the radiation dominated phase \( R_H(t) = 2t \) and matter dominated phases \( R_H(t) = 3t/2 \).

is growing (even exponentially during inflation) whereas the Hubble radius is constant during inflation, this means that at a much earlier time, it has emerged from the Hubble radius of the de Sitter phase. In this scenario, the origin of the fluctuations is thus found in the heart of the de Sitter event horizon: using quantum field theory in curved space, one may compute the amplitude of the quantum fluctuations of the scalar field; their wavelengths evolve as \( a(t)/k \) until they outgrow the event horizon i.e. the Hubble radius; they freeze out and continue to evolve classically. The fluctuation spectrum produced is given by

\[
P_S(k) = \left[ \frac{H^2}{\phi^2} \left( \frac{H}{2\pi} \right)^2 \right]_{k=aH} = \frac{1}{12\pi^2 m_p^6} \left( \frac{V^3}{V''} \right)_{k=aH},
\]

where the subscript \( k = aH \) means that the quantities are evaluated at Hubble radius crossing, as expected. We also note that \( H \) i.e. \( R_H \) sets the scale of quantum fluctuations in the de Sitter phase (see Appendix D for details, in particular Eq. (D.40)): \( R_H \) is indeed the dynamical scale associated with the physics of fluctuations (which happens to coincide with either of the kinematical scales which are the particle horizon for the matter dominated phase, or the event horizon in the inflation phase).

The scalar spectral index \( n_S(k) \) is computed to be (see Exercise D-1 in Appendix D):

\[
n_S(k) - 1 \equiv \frac{d \ln P_S(k)}{d \ln k} = -6\epsilon + 2\eta.
\]

Thus, because of the slow roll, the fluctuation spectrum is almost scale invariant, a result that we have alluded to when we discussed the origin of CMB fluctuations. One of the highlights of the Planck cosmology results \[16, 29\] is the confirmation that the spectrum is not scale invariant i.e. \( n_S \) is different from 1:

\[
n_S = 0.9603 \pm 0.0073.
\]

In other words, we are really in a slow roll phase, i.e. an unstable phase which is crucial since eventually one has to get out of inflation and reheat.

The observation by the COBE satellite of the largest scales has set an important constraint on inflationary models by putting an important constraint on the size of fluctuations (see the caption of Figure 5). Specifically, in terms of the value of the scalar potential at horizon crossing, this constraint known as COBE normalization, reads (see (96):

\[
\frac{1}{m_p^3} \frac{V^{3/2}}{V'} = 5.3 \times 10^{-4}.
\]
Using the slow roll parameter introduced above in (88), the COBE normalization condition can be written as

\[ V^{1/4} \sim 0.03 \varepsilon^{1/4} m_p . \]  

(100)

Besides scalar fluctuations, inflation produces fluctuations which have a tensor structure, i.e. primordial gravitational waves. They can be written as perturbations of the metric of the form

\[ ds^2 = a^2 \left[ \eta_{\mu\nu} + h^{TT}_{\mu\nu} \right] , \]

(101)

where \( h^{TT}_{\mu\nu} \) is a traceless transverse tensor (which has two physical degrees of freedom i.e. two polarizations). The corresponding tensor spectrum is given by

\[ P_T(k) = \frac{8}{m_p^2} \left( \frac{H}{2\pi} \right)^2 , \]

(102)

with a corresponding spectral index

\[ n_T(k) \equiv \frac{d \ln P_T(k)}{d \ln k} = -2\epsilon . \]

(103)

We note that the ratio \( P_T/P_S \) depends only on \( \dot{\phi}^2/H^2 \) and thus on \( \epsilon \), which yields the consistency condition:

\[ r \equiv \frac{P_T}{P_S} = \frac{8\dot{\phi}^2}{m_p^2 H^2} = 16\epsilon = -8n_T . \]

(104)

2.4 Inflation scenarios

We conclude this discussion by reviewing briefly the main classes of inflation models. Let us note that, for an inflationary model, the whole observable universe should be within the Hubble radius at the beginning of inflation. This corresponds to a scale

\[ k = a_0 H_0 = aH|_{h.c.} . \]

(105)

where \( h.c. \) stands for “horizon crossing”. This puts a constraint on the number of e-foldings (93) between horizon crossing and the end of inflation (i.e. end of the slow roll regime) necessary for the inflation to be efficient. More generally, one defines [30] \( N(k) \) as the number of e-foldings between the time of horizon crossing of the scale \( k \) (\( t_{k,h.c.} \) at which \( k = aH(t_{k,h.c.}) \)) and the end of inflation (\( t_e \)):

\[ N(k) \equiv \ln \left( \frac{a(t_e)}{a(t_{k,h.c.})} \right) . \]

(106)

Distinguishing the time when the universe reheats (\( t_{rh} \)) and the time of matter-radiation equality (\( t_{eq} \)), we have

\[
\frac{a(t_{k,h.c.})}{a_0} = e^{-N(k)} \frac{a(t_e)}{a(t_{rh})} \frac{a(t_{eq})}{a(t_{eq})} \frac{a(t_{eq})}{a_0} = e^{-N(k)} \left( \frac{\rho(t_{rh})}{\rho(t_e)} \right)^{1/3} \left( \frac{\rho(t_{eq})}{\rho(t_{rh})} \right)^{1/4} \left( \frac{a(t_{eq})}{a_0} \right) \]

(107)

where we used \( \rho(t_{eq}) = \rho_0 (a_0/a(t_{eq}))^3 \) and we have assumed matter domination between the end of inflation and reheating. Using \( k = aH(t_{k,h.c.}) \) and \( H(t_{k,h.c.})/H_0 = (\rho_{k,h.c.}/\rho_0)^{1/2} \), one obtains from (106), with transparent notations:

\[ N(k) = 62 - \ln \frac{k}{a_0 H_0} - \ln \frac{10^{16} \text{ GeV}}{V_{k,h.c.}^{1/4}} + \ln \frac{V_{k,h.c.}^{1/4}}{V_{e}^{1/4}} - \frac{1}{3} \ln \frac{V_{e}^{1/4}}{\rho_{rh}^{1/4}} . \]

(108)

241
In the rather standard case where $V_{k,h.c.} \sim V_e \sim \rho_{rh} \sim (10^{16} \text{GeV})^4$, this requires 60 e-folding for inflation to be efficient at the scale of the observable Universe. A length scale corresponding to 200 Mpc (hence $k = 2\pi/200$ Mpc) corresponds to $N(k) \sim 50$.

The three main classes of inflation models (see Figs. 10 and 11) are:

- **convex potentials or large field models ($0 < \eta < 2\epsilon$)**
  The potential is typically a single monomial potential:
  \[
  V(\phi) = \lambda m_P^4 \left( \frac{\phi}{m_P} \right)^n,
  \]
  with $n > 1$. Since the slow roll parameters are $\epsilon = n^2 (m_p/\phi)^2/2$ and $\eta = n(n-1)(m_p/\phi)^2$, the slow roll regime corresponds to $\phi \gg m_P \sqrt{n(n-1)}$ (for $n > 2$). Because the field has a value larger than the Planck scale (hence the name “large field model”), this might seem out of the reach of the effective low energy gravitation theory. But A. Linde [20] argued that the criterion is rather $V < m_P^4$, in fact, he suggested that the scalar field emerges from the Planck era with a value $\phi_0$ such that $V(\phi_0) \sim m_P^4$, i.e. $\phi_0 \sim m_P/\lambda^{1/n}$. This corresponds to the chaotic inflation scenario, the simplest example of which being a quadratic potential [20]. A difficulty is that the COBE normalisation imposes an unnaturally small value for the $\lambda$ coupling: $\lambda \sim (5.3 \times 10^{-4}n)^{2n/(n-2)}$. Another drawback is the large value of the field which makes it necessary to include all non renormalisable corrections of order $(\phi/M_p)^{n+p}$, unless they are forbidden by some symmetry.
  The limit case in this class is the exponential potential
  \[
  V(\phi) = V_0 \exp(-\lambda \phi/m_P)
  \]
  which leads to a power law inflation [32]: $a(t) \propto t^{2/\lambda^2}$. This model yields $\eta = 2\epsilon = \lambda^2$, hence $r = -8(n_S - 1)$. It is incomplete since inflation does not end.

- **concave potential or small field models ($\eta < 0$)**
  In this class, illustrated first by the new inflation scenario [19,26], the field $\phi$ starts at a small value and rolls along an almost flat plateau (where $V''(\phi) < 0$) before falling to its ground state. This type of potential, often encountered in symmetry breaking transitions may be parametrized, during

\[\text{Fig. 10: Regions corresponding to the different inflation models in the plot } r \text{ vs } n_s \]
Fig. 11: Constraints set by Planck data on various inflation models in the plot $r$ (evaluated at the pivot scale $k_\ast = 0.002 \text{ Mpc}^{-1}$) vs $n_s$ [29]. Small dots correspond to models with 50 e-foldings, large dots to 60 e-foldings.

the phase transition by:

$$V(\phi) = V_0 \left[1 - \left(\frac{\phi}{\mu}\right)^p + \cdots\right], \quad (111)$$

where the dots indicate higher order terms not relevant for inflation. In the same class appears the so-called “natural inflation” potential [33, 34]

$$V(\phi) = V_0 [1 + \cos(\phi/f)] . \quad (112)$$

A difficulty shared by the class of small field models is the unnaturalness of the initial conditions: why start at the height of the potential, in a plateau region or close to an unstable extremum? In the case of a symmetry breaking potential, the rationale could be thermal: the restoration of the symmetry at high temperature naturally leads to start at the unstable “false vacuum”.

The first model proposed for inflation [4] was based on a modification of gravity described by the following action:

$$S = \frac{m_p^2}{2} \int d^4x \sqrt{-\tilde{g}} \left(\tilde{R} + \alpha \tilde{R}^2\right) , \quad (113)$$

where $\tilde{R}$ is the Ricci scalar associated with the metric $\tilde{g}_{\mu\nu}$. It may be proved that this is equivalent to standard Einstein gravity plus a scalar field with a potential that falls in the same class as we just discussed (see Exercise 2-1).

– hybrid models ($0 < 2\epsilon < \eta$)

The field rolls down to a minimum of large vacuum energy (where $V''(\phi) < 0$) from a small initial value. Inflation ends because, close to this minimum, another direction in field space takes over and brings the system to a minimum of vanishing energy [27]. Such models, which thus require several fields, were constructed in order to allow inflation at scales much smaller than the Planck scale. In this case (see (100), $\epsilon$ can be very small and thus $r$ as well. This class of model is the only one that can accommodate values of $n_S$ larger than 1 (i.e. a blue spectrum).

We note that, for most models, the inflation scale is much larger than the TeV scale. This should in principle lead us to consider inflation models in the context of supersymmetry, in order to avoid an undesirable fine tuning of parameters. In fact, supersymmetric (and superstring) theories are plagued with the presence of numerous flat directions: this might be a blessing for the search of
inflation potential\textsuperscript{14}. Two such types of potentials rely on the properties of basic supersymmetry multiplets: they are called $F$-term \textsuperscript{35,36} and $D$-term \textsuperscript{37,38} inflation and fall in the category of hybrid inflation. They are not presently favoured by Planck data because they give too large values of $n_S$ (of the order of 0.98).

**Exercise 2-1**: We show that a certain class of 4-dimensional models which extend Einstein theory are equivalent to Einstein gravity with a scalar field coupled to the metric (i.e. a scalar-tensor theory). This class of models is described by the following action:

\[
S = \frac{m_p^2}{2} \int d^4x \sqrt{-g} \left[ f(\tilde{R}) + S_m(\psi, \tilde{g}_{\mu\nu}) \right],
\]

where $\tilde{R}$ is the Ricci scalar associated with the metric $\tilde{g}_{\mu\nu}$, and $f(\tilde{R})$ is a general function of this Ricci scalar. We note that the Starobinsky action (113) corresponds to $f(\tilde{R}) = \tilde{R} + \alpha \tilde{R}^2$.

Let us consider the more general action:

\[
S = \frac{m_p^2}{2} \int d^4x \sqrt{-g} \left[ \frac{df}{d\chi}(\tilde{R} - \chi) + S_m(\psi, \tilde{g}_{\mu\nu}) \right].
\]

a) Show that the variation of (116) with respect to $\chi$ leads to (115).

b) Redefining the metric and the scalar field through

\[
g_{\mu\nu} \equiv \frac{df}{d\chi} \tilde{g}_{\mu\nu} , \quad \phi \equiv -\sqrt{\frac{3}{2}} m_p \log \frac{df}{d\chi},
\]

show that one recovers the familiar form of the scalar-tensor gravity:

\[
S = \int d^4x \sqrt{-g} \left[ \frac{m_p^2}{2} \tilde{R} - \frac{1}{2} \partial^\mu \phi \partial_\mu \phi - V(\phi) \right] + S_m(\psi, A^2(\phi) g_{\mu\nu}),
\]

where one will express the potential $V(\phi)$ in terms of $\chi$, $f(\chi)$ and $df/d\chi$, and give the explicit form of $f(\phi)$.

c) Identify the potential $V(\phi)$ in terms of $\chi$, $f(\chi)$ and $df/d\chi$, and give the explicit form of $f(\phi)$.

**Hints**: a) $\chi = \tilde{R}$, under the condition that $f'' \neq 0$.

b) $V \equiv m_p^2 \frac{\chi df/d\chi - f}{2(df/d\chi)^2}$ and $A(\phi) = \exp \left[ \phi / (m_p \sqrt{6}) \right]$.

3 Light does not say it all (1): the violent Universe

The Universe is the siege of many violent phenomena; one may cite explosions like supernovae, gamma ray bursts (GRB) or the emission of energetic particles by active galaxy nuclei (AGN), quasars, blazars... The time constants $\tau$ associated with the phenomena are very short on the scale of the Universe. For example, a GRB may be visible on the sky only for a few seconds. This means that the distance scales $c \tau$ involved are very small: the distance that light travels in 10 seconds is only 3 million km, that is 0.002 astronomical unit (1 a.u. is the Sun-Earth distance). Indeed, very compact objects, such as neutron stars or black holes, are at the heart of such violent phenomena. We will start by reviewing the origin of such compact astrophysical objects, which appear at the end of the life of a star.

\textsuperscript{14}One possible difficulty arises from the condition (89) which may be written as a condition on the mass of the inflaton field

\[
m^2 \ll H^2.
\]

Any fundamental theory with a single dimensionful scale (such as string theory) runs into the danger of having to fine tune parameters in order to satisfy this constraint. This is known as the $\eta$ problem.
3.1 The end of the life of a star: from white dwarfs to neutron stars and black holes

The evolution of a generic gravitational system such as a star is governed by two competing processes: gravitational forces which tend to contract the system and thermal pressure which is due to the thermonuclear reactions within, which tend to expand the system. In a stable star like our Sun at present, the two processes balance each other. But when the nuclear fuel is exhausted, the (core of the) star starts to collapse under the effect of gravity; the gravitational energy thus released heats up the outer layers of the star, which produces the explosive phenomena that we observe.

But what is the fate of the collapsing core? Gravitational pressure is eventually counterbalanced by quantum degeneracy pressure. Let us explain the nature of this pressure. Since matter is made of fermions of spin 1/2, Pauli principle applies: two fermions cannot be in the same state. Fermionic matter will thus resist at some point to excessive pressure.

Let us be more quantitative. Since there are $4\pi\rho^2 dp/(2\pi\hbar)^3$ levels per unit volume with momentum between $p$ and $p+dp$ and two spin states per level, the number of fermions per unit volume is given in terms of the maximal momentum by

$$n = \frac{2}{(2\pi\hbar)^3} \int_0^{p_F} 4\pi k^2 dk = \frac{p_F^3}{3\pi^2\hbar^3}. \tag{119}$$

The energy of the highest level, or Fermi energy $\epsilon_F$, is therefore given in terms of the number density $n$. If the particles are non-relativistic, then

$$\epsilon_F = \frac{p_F^2}{2m} = \frac{1}{2} \left(\frac{3\pi^2}{2}\right)^{2/3} \hbar^2 n^{2/3} m. \tag{120}$$

On the other hand, the gravitational energy per nucleon of a system of size $R$ and mass $M$ (with $N = 4\pi R^3 n_N/3 = M/m_N$ nucleons) is

$$\epsilon_g = \frac{G_N M m_N}{R} = G_N m_N^2 \frac{N}{R} = \left(\frac{4\pi}{3}\right)^{1/3} G_N m_N^2 N^{2/3} n_N^{1/3}. \tag{121}$$

The Fermi energy starts to dominate over the gravitational energy for

$$n_N^{1/3} > \frac{2}{(3\pi^2)^{2/3}} \left(\frac{G_N m_N^2 m}{\hbar^2}\right) N^{2/3} \nu^{2/3}, \tag{122}$$

where $\nu = n_N/n$ ($\nu$ depends on the species of the fermions that are degenerate; see below), or

$$RM^{1/3} < \frac{1}{\alpha_G} \frac{\hbar}{mc} m_N^{1/3} \nu^{-2/3}, \tag{123}$$

where, as above, $\alpha_G \equiv (G_N m_N^2/\hbar c) \sim 6 \times 10^{-39}$.

We see from (122) that gravitational collapse is first stopped by the quantum degeneracy of electrons: the corresponding astrophysical objects are known as white dwarfs. Writing thus $m = m_e$ and $\nu = 2$ (two nucleons per electron), we find that $RM^{1/3} \sim 10^{-2} R_\odot M_\odot^{1/3}$. A white dwarf with $M = M_\odot$ has radius $R \sim 10^{-2} R_\odot$ and density $\rho \sim 10^6 \rho_\odot$. It is more compact than a star.

If density continues to increase, the value of the Fermi energy is such that the fermions are relativistic: it follows from (119) that $p_F > mc$ reads $\left(\frac{3\pi^2}{2}\right)^{1/3} \hbar n^{1/3} > mc$ or, using $n = 3N/(4\pi\nu R^3)$,

$$R < \left(\frac{9\pi}{4}\right)^{1/3} \frac{\hbar c}{mc^2} N^{1/3} \nu^{-1/3}. \tag{124}$$
But, since $\epsilon_F \sim p_{Fc} = (3\pi^2)^{1/3} \hbar n^{1/3}$, both $\epsilon_F$ and $\epsilon_g$ scale like $n^{1/3}$. Quantum degeneracy pressure can overcome gravitational collapse only for $N < 3\sqrt{\pi} \alpha_G^{-3/2} / (2\nu^2)$, or

$$M < 3\sqrt{\pi} \alpha_G^{-3/2} m_N / (2\nu^2) \sim 1 \ M_\odot / \nu^2.$$  

(125)

This bound is the well-known Chandrasekhar limit for white dwarf masses (a more careful computation gives a numerical factor of 5.87 [39]). The radius of the object then satisfies (see (124))

$$R < \frac{3\sqrt{\pi}}{2} \alpha_G^{-1/2} \frac{\hbar e}{m c^2 \nu}.$$  

(126)

Setting $m = m_e$ gives a limit value of some $10^4$ km.

For even higher densities, most electrons and protons are converted into neutrons through inverse beta decay ($p + e^- \rightarrow n + \nu$). A new object called neutron star forms when the neutron Fermi energy balances the gravitational energy. Writing $m = m_n$ instead of $m_e$ in (123), we now have $RM^{1/3} \sim 10^{-5} R_\odot M_\odot^{1/3}$: a neutron star with $M = M_\odot$ has radius $R \sim 10^{-5} R_\odot$ and density $\rho \sim 10^{15} \rho_\odot$.

The bound (125) obtained above in the case of relativistic fermions (neutrons in this case) is called the Oppenheimer-Volkoff bound: more precisely, the maximal mass of a neutron star is $M = 0.7 \ M_\odot$, with a corresponding radius $R = 9.6$ km (cf. (126) with $m = m_n$). If the mass is larger, the star undergoes gravitational collapse and forms a black hole.

### 3.2 Gravitational collapse: black holes

Let us first backtrack a little and return to Einstein’s equations (3). Because they are non-linear, there are few solutions known. The first exact non-trivial solution was found in late 1915 by Schwarzschild, who was then fighting in the German army, within a month of the publication of Einstein’s theory and presented on his behalf by Einstein at the Prussian Academy in the first days of 1916 [9], just before Schwarzschild death from a illness contracted at the front. It describes static isotropic regions of empty spacetime, such as the ones encountered in the exterior of a static star of mass $M$ and radius $R$.

The Schwarzschild solution reads, for $r > R$ (see Exercise 3-1),

$$ds^2 = \left( 1 - \frac{2M_N}{r} \right) dt^2 - \left( 1 - \frac{2G_N M}{r} \right)^{-1} dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2.$$  

(127)

The Schwarzschild solution is singular at $r = R_S \equiv 2G_N M$, a distance known as the Schwarzschild radius. This is not a problem as long as $R_S < R$ since this solution describes the exterior region of the star. A different metric describes the interior. On the other hand, we will see in Section 3.2.2 that, in the case where $R < R_S$, i.e. $2G_N M / R > 1$, the system undergoes gravitational collapse and turns into a black hole.

**Exercise 3-1**: In this exercise, we derive the Schwarzschild solution (127). Because we look for static isotropic solutions, we may always write the spacetime metric as\[^{15}\]:

$$ds^2 = e^{2\nu(r)} dt^2 - e^{2\lambda(r)} dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\phi^2).$$  

(128)

In other words, the only non-vanishing elements of the metric are:

$$g_{tt} = e^{2\nu(r)}, \ g_{rr} = -e^{2\lambda(r)}, \ g_{\theta\theta} = -r^2, \ g_{\phi\phi} = -r^2 \sin^2 \theta.$$  

(129)

\[^{15}\]We have absorbed a general function $e^{2\nu(r)}$ in front of the last term by redefining the variable $r$.

\[\text{246}\]
b) Show that the Einstein’s equations in the vacuum simply amount to a condition of vanishing Ricci tensor:

\[ R_{\mu\nu} = 0 . \]  

(130)

c) From the vanishing of \( R_{tt} \) and \( R_{rr} \) and the fact that, at large distance from the star, space should be flat, both \( \lambda \) and \( \nu \) should vanish at spatial infinity. Hence deduce that

\[ \lambda = -\nu . \]  

(131)

d) From the vanishing of \( R_{\theta\theta} \), deduce that

\[ g_{tt} e^{2\nu} = 1 - \frac{2G_N M}{r} . \]  

(132)

Hints: a)

\[
R_{tt} = \left( \nu'' + \nu'^2 - \lambda' \nu' + \frac{2\nu'}{r} \right) e^{2(\nu-\lambda)} ,
\]

\[
R_{rr} = -\nu'' - \nu'^2 + \lambda' \nu' + \frac{2\lambda'}{r} ,
\]

\[
R_{\theta\theta} = 1 - (1 + \nu'/r) \lambda e^{-2\lambda} ,
\]

\[
R_{\phi\phi} = R_{\theta\theta} \sin^2\theta .
\]

(133)

b) (3) reads \( R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 0 \). Contracting with \( g^{\alpha\mu} \) yields \( R = 0 \).

d) The constant of integration is identified with the mass \( M \) because, in the Newtonian limit, \( g_{tt} = 1 + 2\Phi \) where \( \Phi \) is the Newtonian potential.

Exercise 3-2: What is the Schwarzschild radius of the sun? of an astrophysical object of mass \( 3 \times 10^6 M_{\odot} \)?

Hints: Do not forget that we have set \( c = 1 \). Otherwise, \( R_S = 2G_N M/c^2 \), that is 2.95 km for the sun, \( 8.85 \times 10^9 \) m = 0.06 au for an object of mass \( 3 \times 10^6 M_{\odot} \).

We now understand that, when (the core of) a star of mass \( M \) in gravitational collapse overcomes the degeneracy pressure of neutrons to reach a size \( R < R_S = 2G_N M \), nothing seems to drastically change for observers located at distances \( r > R_S \). However, the behaviour of the Schwarzschild metric 127 appears to be singular: \( g_{tt} \) vanishes and \( g_{rr} \) diverges. It took some time (Lemaître again!) to realize that this was not the sign of a real singularity but was just an artifact of the choice of coordinates: other choices lead to a regular behaviour (see Exercise 3-3). The true singularity lies at \( r = 0 \) where the collapsing matter ends up.

In order to understand the nature of the surface at \( r = R_S \), let us keep for a moment longer the Schwarzschild coordinates and consider sending a light signal radially from some point \( r_1 \) to \( r_2 > r_1 \) where it is received a time \( \Delta t \) later. Since \( ds^2 = 0 \) (as well as \( d\theta = d\phi = 0 \)), we have simply

\[ \Delta t = \int_{r_1}^{r_2} \frac{dr}{(1 - R_S/r)} . \]  

(134)

If \( r_1 < R_S \), this is finite only for \( r_2 < R_S \), in which case it is simply \( r_2 - r_1 + R_S \ln [(R_S - r_2)/(R_S - r_1)] \). In other words, signals emitted from within the Schwarzschild radius never reach the outside. There is really a breach of communication. Indeed, the surface \( r = R_S \) is an event horizon (see Section 2.1).
Let us take this opportunity to present a classical interpretation of the Schwarzschild radius. Remember that the existence of black holes was conceived by Michell [40] and Laplace [41] centuries earlier than general relativity. Indeed, the classical condition for escape a body of mass \( m \) and velocity \( v \) from a spherical star of mass \( M \) and radius \( R \) is

\[
\frac{1}{2}mv^2 > \frac{G_NM}{R}.
\]

Thus, not even light \((v = c)\) can escape the attraction of the star if \( R < 2G_NM/c^2 \), the Schwarzschild radius.

We note that the Schwarzschild horizon is a fictitious surface, in the sense that an observer crossing this surface would not experience anything particular (we said that there exist coordinates where the behaviour at \( R_S \) is regular), except deformations due to tidal forces because it comes closer to a very massive object. But once it has crossed this fictitious surface, there is no way to backtrack: the further information that might be gained is lost for ever to the outside world. A useful picture is the one of a person swimming in a river with a waterfall downstream: swimming in the river involves no danger as long as one is safely far from the waterfall, but, at some point the swimmer crosses a fictitious line (the “horizon” of the waterfall) which is the point of no return: even the best swimmer is attracted towards the “singularity” of the waterfall.

So far, our description has been purely classical. Quantum mechanical processes change this picture. Indeed, S. Hawking [42] pointed out that black holes emit radiation through what is known as the process of evaporation. Indeed, it can be shown that an accelerated observer sees a thermal bath of particles at a temperature measured by an observer at infinity to be

\[
T_H = \frac{\hbar}{4\pi R_S}.
\]

The Hawking evaporation process is important to understand the non-observation of primordial black holes, which would be due to fluctuations of density during the Planck era: such primordial black holes have evaporated.

A final comment using the Schwarzschild coordinates (127): we see that, when \( r \) crosses \( R_S \), the respective signs of \( g_{tt} \) and \( g_{rr} \) changes. In other words, \( t \) becomes a spatial coordinates and \( r \) becomes time: the movement towards the central singularity is the clock that ticks.

**Exercise 3-3**: Define the Kruskal coordinates \((v, u, \theta, \phi)\) related to the Schwarzschild coordinates \((t, r, \theta, \phi)\) through [44]:

\[
\begin{align*}
\text{for } r > R_S, & \quad u = (r/R_S - 1)^{1/2} e^{r/2R_S} \cosh(t/2R_S), \\
& \quad v = (r/R_S - 1)^{1/2} e^{r/2R_S} \sinh(t/2R_S), \\
\text{for } r < R_S, & \quad u = (1 - r/R_S)^{1/2} e^{r/2R_S} \sinh(t/2R_S), \\
& \quad v = (1 - r/R_S)^{1/2} e^{r/2R_S} \cosh(t/2R_S).
\end{align*}
\]

Deduce from (127) the form of the metric in Kruskal coordinates:

\[
ds^2 = \frac{4R_S^3}{r} e^{-r/R_S} \left( dv^2 - du^2 \right) - r^2 \left( d\theta^2 + \sin^2 \theta d\phi^2 \right),
\]

where \( r \) is given as an implicit function of \( u \) and \( v \):

\[
\left( \frac{r}{R_S} - 1 \right) e^{r/R_S} = u^2 - v^2.
\]
In order to be more quantitative, let us follow the analysis of Oppenheimer and Snyder [45] who were the first to discuss the collapse into a black hole. We consider a fluid of negligible pressure, thus described by the energy-momentum tensor (see (21)) \( T_{\mu\nu} = \rho U_\mu U_\nu \), and study its spherically symmetric collapse.

It turns out that we have already studied this system when we discussed the evolution of a homogeneous and isotropic universe in Section 1.3. The metric is given by

\[
 ds^2 = dt^2 - a^2(\hat{t}) \left( \frac{d\hat{r}^2}{1 - k\hat{r}^2} + \hat{r}^2 \, d\hat{\theta}^2 + \hat{r}^2 \sin^2 \hat{\theta} \, d\hat{\phi}^2 \right)
\]

(140)
as in (19)\textsuperscript{16} and the Einstein tensor components are the same as in (B.5,B.6). We normalize the coordinate \( \hat{r} \) so that \( a(0) = 1 \). Thus

\[
 \rho(\hat{t}) = \rho(0)/a^3(\hat{t})
\]

(141)

and Einstein’s equations simply read:

\[
 a^2 + k = \frac{8\pi G N}{3} \rho(0)/a \ ,
\]

(142)

\[
 \dot{a}^2 + 2a\ddot{a} + k = 0 \ .
\]

(143)

Assuming that the fluid is initially at rest (\( \dot{a} = 0 \)), we obtain from (142)

\[
 k = \frac{8\pi G N}{3} \rho(0) \ .
\]

(144)

Thus, (142) simply reads

\[
 \dot{a}^2(\hat{t}) = k \left[ a^{-1}(\hat{t}) - 1 \right] .
\]

(145)

The solution is given by the parametric equation of a cycloid:

\[
 \begin{align*}
 \hat{t} &= \frac{\psi + \sin \psi}{2\sqrt{k}} , \\
a &= 1 + \cos \psi \\
\end{align*}
\]

(146)

We see that \( \dot{a} \) vanishes for \( \psi = \pi \), that is after a time

\[
 \tau = \frac{\pi}{2\sqrt{k}} = \frac{\pi}{2} \left( \frac{3}{8\pi G N \rho(0)} \right)^{1/2} .
\]

(147)

Thus a sphere initially at rest with energy density \( \rho(0) \) and negligible pressure collapses to a state of infinite energy density in a finite time \( \tau \).

In the case of a star of radius \( R \) and mass \( M \), this solution for the interior of the star should be matched with the Schwarzschild solution (127) describing the exterior. The correspondence between the interior and exterior coordinates is simply \( r = Ra(\hat{t}), \theta = \hat{\theta} \) and \( \phi = \hat{\phi} \), with a more complex relation between \( t \) and \( \hat{t} \) (see Ref. [39] section 11.9). The first relation ensures that

\[
 k = \frac{2MG_N}{R^3} = \frac{R_S}{R^3} ,
\]

(148)
in agreement with (144) and \( M = (4\pi/3)\rho(0)R^3 \).

\textsuperscript{16} except that we do not normalize \( k \) to \( \pm 1 \) or \( 0 \) because we are looking at a different system. We will see just below that it is fixed by initial conditions. We add a hat to this system of coordinates to distinguish it from the Robertson-Walker coordinates, as well as from the Schwarzschild coordinates that we will use later.
Exercise 3-4: We consider an observer at rest outside the horizon of a black hole described by the Schwarzschild metric (127) which we write (see Exercise 3-1):

\[ ds^2 = e^{2\nu(r)} dt^2 - e^{-2\nu(r)} dr^2 - r^2 \left( d\theta^2 + \sin^2 \theta d\phi^2 \right). \]

The observer velocity is \( U^\mu = \xi^\mu / \sqrt{\xi^2} \) with \( \xi^\mu \equiv \delta_0^\mu \).

a) Show that \( \xi_{\mu;\nu} = \xi_{\mu} w_{;\nu} - \xi_{\nu} w_{;\mu} \), where \( w_{;\mu} \equiv \partial_{;\mu} \xi \) (the covariant derivative \( \xi_{\mu;\nu} \) is defined in (B.4) of Appendix B).

b) Deduce that the acceleration \( A^\mu \equiv U^\rho \nabla_\rho U^\mu \) of the observer is simply \( A^\mu = -w^\mu \).

c) Show that the acceleration \( a^2 \equiv -w^\mu w_{;\mu} \) is given for the observer at fixed \( r, \theta \) and \( \phi \) by

\[ a = \frac{G_N M}{r^2 \left( 1 - 2G_N M/r \right)^{3/2}}. \]

We finally note that, at large distance, the black hole is only characterized by its mass \( M \). Black holes are indeed very simple objects, somewhat similar to particles: Schwarzschild black holes are only characterized by their mass. Other more complex solutions were found later but it was realized that one can only add spin (rotating or Kerr black holes) and charge (charged black holes) but no other independent characteristics: in the picturesque language used by Wheeler, it is said that black holes can have no hair. In a sense, the black holes of general relativity are very similar to fundamental particles, which are characterized by a finite set of numbers (including mass, spin, and electric charge).

Astrophysical black holes are somewhat more complex because of their material environment, as we will now see.

3.3 Astrophysical black holes

For a long time, black holes were considered as a curiosity of general relativity and did not have the status of other stellar objects. This has changed in the last decade which has seen mounting evidence that, at the center of our own galaxy (Milky Way) cluster, there is a massive black hole associated with the compact radio source Sagittarius A*. Observations of the motions of nearby stars by the imager/spectrometer NAOS/CONICA working in the infrared [46] have indeed confirmed the presence of a very massive object \((2.6 \pm 0.2).10^6\) solar mass) localized in a very small region (a fraction of an astronomical unit, see Fig. 12), which seems only compatible with a black hole.

Since then, black holes have been identified in many instances and their role might be central in many phenomena. We have seen that they are very simple gravitational objects. But these simple objects accrete matter, and are thus associated with very diverse phenomena. A picture has emerged, which seems to be valid at very diverse scales (see Fig. 13) of a black hole surrounded by an accretion disk and a torus of dust, with two opposite relativistic jets, which are supposedly formed during the gravitational collapse through a recombination of the magnetic fields. Of course, at the centre of this complex structure, lies the black hole surrounded by its horizon. But the complex phenomena that take place in this surrounding region allow to detect indirectly the black hole.

Let us review some of the astrophysical occurrences of black holes.

First, our galaxy is not the only one which has a central black hole. This is believed to be very general, and in many cases the black hole and its environment is much more active than our own. Active galaxies are galaxies where the dominant energy output is not due to stars. In the case of Active Galactic Nuclei (AGN), the non-thermal radiation comes from a central region of a few parsecs around the centre of the galaxy. The most famous example of such AGNs is provided by quasi-stellar objects (QSOs) or quasars: these starlike objects turn out to be associated with the point-like optical emission from the nucleus of an active galaxy.
The typology of active extragalactic objects is very complex: radio loud and radio quiet quasars, Seyfert galaxies, BL Lacs or blazars... There has been an effort to build a unified picture [47]: the apparent diversity in the observations would then result from the diversity of perspectives from which we observers see these highly non-isotropic objects. Typically, the model for radio-loud AGNs includes (see Fig. 13, right panel): a central engine, a pair of oppositely directed relativistic jets (cones of semi-angle around $1\degree$) an accretion disk (of size of the order of 1 parsec), and a torus of material (of size of the order of 100 parsec) which obscures the central engine when one observes it sideways. Depending on the relative angle between the line of sight and the jet axis, observation may vary in important ways.

Gamma ray bursts (GRB) are the most luminous events observed in the universe. They were discovered accidentally by the American military satellites VEGA which were designed to monitor the nuclear test ban treaty of 1963. The first burst was found in 1969, buried in gamma-ray data from 1967: two Vela satellites had detected more or less identical signals, showing the source to be roughly the same distance from each satellite [48].

A GRB explosion can be as luminous as objects which are in our vicinity, such as the Crab nebula, although they are very distant. The initial flash is short (from a few seconds to a few hundred for a long GRB, a fraction of a second for a short one). From 1991 to 2000, BATSE (Burst and Transient Source Experiment) has allowed to detect some 2700 bursts and showed that their distribution is isotropic, a good argument in favor of their cosmological origin. In 1997 (February 28), the precise determination of the position of a GRB (hence named 970228) by the Beppo-SAX satellite allowed ground telescopes to discover a rapidly decreasing optical counterpart, called afterglow. Typically in the afterglow, the photon energy decreases with time as a power law (from X ray to optical, IR and radio) as well as the flux: it stops after a few days or weeks. The study of afterglows gives precious information on the dynamics of GRBs. The launch of the SWIFT satellite on 20 November 2004 has started a new era for the understanding of GRBs.

Given the time scales involved (a few milliseconds for the rise time of the gamma signal), the size of the source must be very small: it cannot exceed the distance that radiation can travel in the same time interval, i.e. at most a few hundred kilometers. Energy must have been ejected in an ultra-relativistic flow which converted its kinetic energy into radiation away from the source: the Lorentz factors involved are typically of the order of 100!

This flow is collimated and forms a jet of half opening angle $\theta$. The observational evidence for
this collimation is an achromatic break in the afterglow light curve: for \( t > t_{\text{jet}} \), it decreases faster than it would in the spherical case \([49,50]\). If we assume that the relativistic jet, after emitting a fraction \( \eta_\gamma \) of its kinetic energy into prompt \( \gamma \) rays, hits a homogeneous medium with a constant number density \( n \), the break appears in the afterglow light curve when the Lorentz factor \( \gamma \) becomes of the order of \( 1/\theta \). This gives a relation between the half opening angle \( \theta \) and the break time \( t_{\text{jet}} \) \([50]\):

\[
\theta = \frac{0.161}{(1 + z)^{3/8}} \left( \frac{t_{\text{jet}}}{1 \text{ day}} \right)^{3/8} \left( \frac{10^{52} \text{ ergs}}{E_{\gamma,\text{iso}}} \right)^{1/8} \left( \frac{n}{1 \text{ cm}^{-3}} \right)^{1/8} \eta_\gamma^{1/8},
\]

where \( E_{\gamma,\text{iso}} \) is the isotropic equivalent gamma ray energy.

In the collapsar model of Woosley \([51]\), long GRBs are associated with the explosion of a rapidly rotating massive star which collapses into a spinning black hole (see Fig. 13 left panel). The burst and its afterglow have been successfully explained by the interaction of a highly relativistic jet with itself (internal shocks \([52, 53]\)) and with the circumstellar medium (external shocks \([54]\)). Typically, one expects per day \( 10^6 \) collapses of massive stars in the Universe; \( 10^3 \) give rise to a GRB and approximately 1 of these is pointing towards us its jet. Hence, one may observe from earth about one GRB per day.

Supernovae explosions also provide very bright events in the sky, some of them being visible to the naked eye. Supernovae explosions were thus recorded in 1006, 1054, 1181, 1572 and 1604. The Sn 1987 A explosion allowed the detection of neutrinos and gamma emission.

---

Fig. 13: Unified picture of the system surrounding an astrophysical black hole in the case of a gamma ray burst (left), a microblazar (centre) and a blazar (right). In the first two cases, the central object is a stellar mass black hole and the size of the accretion disk is about 1000 km. In the latter (AGN), it is a supermassive black hole of potentially several million solar masses and the accretion disk is of the order of 1 billion kms.
The modern theory of supernovae was initiated in the 30s by Baade and Zwicky [55].

Supernovae follow a classification according to spectroscopy. In type I supernovae, hydrogen lines are absent whereas they are present in type II. Moreover, type I has subclasses: for example, type Ia involves intermediate mass elements (Si). Each type corresponds to a different mechanism for the explosion. In particular, type II and type Ia have a completely different interpretation.

We will focus first on type II supernovae.

Presupernova stars ($M > 8M_\odot$) have an onion-like structure. From the outer to the inner layers, one finds increasingly heavy elements: $H$, $He$, $C$, $O$, $Ne$, $Si$ and $Fe$.

As $Si$ is consumed by nuclear reactions, the mass of the $Fe$ core increases. The resulting density increase then turns the electrons relativistic and makes electronic capture ($p + e \rightarrow n + \nu$) energetically favorable. This diminishes the degenerate electron pressure and leads to the collapse of the core. Since $\rho_{core} \sim 10^{12}$ kg.m$^{-3}$, the collapse time is typically $(G_N\rho)^{-1/2} \sim 0.1$ s.

This time, neutrinos produced as electrons are turned into neutrons, are trapped in the imploding core. The critical density for which neutrinos are trapped, is typically $\rho \sim 2 \times 10^{14}$ kg/m$^3$. As the core is crushed to higher densities, the density approaches that of a neutron star ($\rho \sim 2 \times 10^{17}$ kg/m$^3$) and matter becomes almost incompressible. If the process was elastic, the kinetic energy would be enough to bring it back to the initial state. Typically

$$E \sim G_N M_{core} \left( \frac{1}{R_{NS}} - \frac{1}{R_{WD}} \right) \sim \frac{G_N M_{core}}{R_{NS}} \sim 3 \times 10^{46}$$. \hspace{1cm} (152)

This is not completely so but there is a rebound of the core which sends a shock wave outward. Meanwhile, the stellar matter has started to free fall since it is no longer sustained by its core. The falling matter meets the outgoing shock wave and turns it into an accretion wave.

Neutrinos emitted from the core heat up and expand the bubble thus formed. Convection and neutrino heating thus convey a fraction of the order of one percent of the neutron star gravitational mass (152) to the accretion front. This is enough to make it explode.

One word of caution however: numerical models that try to reproduce supernovae explosions have been until now unable to explode the supernova! One needs to start the explosion artificially. It therefore remains possible that one is still missing a key ingredient in the recipe.

The bulk of the star blown off by the explosion makes what is known as a supernova remnant. It sweeps the interstellar medium at great velocity ($10000$ km/s) and may remain visible for $10^5$ to $10^6$ years. A large fraction of the interstellar medium is thus swept by supernovae remnants (see Exercise 3-1). This is important since this is believed to be the way the heaviest nuclear elements are scattered in the universe (primordial nucleosynthesis produces no element heavier than $^7Li$).

**Exercise 3-4:** a) Assuming approximately one supernova explosion every 30 years in our galaxy (assimilated to a disk of radius $15$ kpc and thickness $200$ pc), compute the corresponding rate $\mathcal{R}$ of supernovae explosions per pc$^3$ and per year.

b) If every supernova leads to a remnant of radius $R = 100$ pc that lasts for $t \sim 10^6$ yrs, what fraction of the galaxy volume is filled by the supernova remnant?

Hints: a) $\mathcal{R} \sim 2.3 \times 10^{-13}$ pc$^{-3}$ yr$^{-1}$.

b) $1 - \exp \left[ -\left(4\pi/3\right) R^3 \mathcal{R} t \right] \sim 0.5$.

SNIa events on the other hand are thermonuclear explosions of white dwarfs. More precisely, a carbon-oxygen white dwarf accretes matter (from a companion star or by coalescence with another white dwarf) which causes its mass to exceed the Chandrasekhar limit. The central core collapses, making the
carbon burn and causing a wave of combustion to propagate through the star, disrupting it completely. The total production of energy is thus almost constant. For a white dwarf of radius 1500 to 2000 km, about $2 \times 10^{51}$ ergs is released in a few seconds during which takes place the acceleration of the material. This is followed by a period of free expansion. Virtually all the energy of the explosion goes into the expansion. The luminosity of the supernova, on the other hand, finds its origin in the nuclear decay of the $^{56}$Ni freshly synthesized. The energy release in the nuclear decays $^{56}$Ni $\rightarrow$ $^{56}$Co $\rightarrow$ $^{56}$Fe, with respective lifetimes of 8.8 and 111 days, represents a few percent of the initial energy release.

This model allows to understand the homogeneity of the observed type Ia supernovae explosions and why they have been used successfully as standard candles in cosmology (see Section 5.1 of Chapter 5). The structure of a white dwarf is determined by degenerate electrons and thus independent of detailed chemical composition (see Section 3.1). The rate of expansion is set by the total energy available since the complete white dwarf is disrupted. Finally, the absolute brightness is determined by the radioactive decay of $^{56}$Ni produced during the explosion. Less Ni means a lower luminosity but also lower temperature in the gas and thus lower opacity and more rapid energy escape. Thus dimmer supernovae are quicker i.e. have narrower light curves.

### 3.4 High energy cosmic particles

As explained above, compact objects and the violent phenomena associated with their formation are important to understand the origin of high energy cosmic particles. It is important to identify the potential sites of acceleration. Obviously, the jets described in the preceding Section are sources of energetic particles. Shock fronts, such as supernova remnants, are also the siege of acceleration for particles whose multiple scattering off magnetized clouds lead to multiple encounters with the shock front.

M. Hillas [56] has proposed a general discussion of potential acceleration sites, in terms of the magnetic fields $B$ available and the size $R$ of the site. The Larmor radius of the particle $\tau_L = E/(qBc)$ (in relativistic regime) may, with increasing energy $E$, become larger than the dimension $R$ of the accelerating site. We thus have the condition ($q = Ze$)

$$E < E_{\text{max}} = qBcR = Z \left( \frac{B}{1 \mu G} \right) \left( \frac{R}{1 \text{ Mpc}} \right) 9.3 \times 10^{20} \text{ eV}. $$

(153)

Note that this is the work of the electric field $E = Bc$ over the maximal distance $R$. In the case of acceleration on magnetic clouds or a shock wave (where $E = BV$), the maximal energy reads:

$$E_{\text{max}} = ZeBV R. $$

(154)

In the case where acceleration involves large Lorentz factors $\gamma$, an extra factor $\gamma$ should be included to account for the energy being measured in the lab frame (also $E = \gamma BV$).

This general criterion allows to draw the now classical Hillas diagram which identifies the possible acceleration sites in a plot $\log (B/1 \text{ G})$ vs. $\log (R/1 \text{ km})$. As can be seen on Fig. 14, given species of cosmic particles accelerated at given energies are represented by diagonal lines (from top to bottom on the figure: protons of $10^{21} \text{ eV}$, protons of $10^{20} \text{ eV}$ and iron nuclei of $10^{20} \text{ eV}$).

### 4 Light does not say it all (2): dark matter

We have known for a long time that the Universe has a dark component, and is not only luminous matter and radiation: already in 1933, by studying the velocity distribution of galaxies, Zwicky [57] discovered that the Coma cluster had 400 times more mass than expected from its luminosity. Through the XXth century, it was realized that non-luminous matter, dark matter, is needed at all scales from galactic to cosmological. This pleads for a form of matter which is not included in the Standard Model, hence for physics beyond the Standard Model. Indeed, the detection of dark matter might be the first sign of physics beyond the Standard Model. Hence the programs of direct or indirect dark matter detection are of key importance not only for astrophysics but also for high energy physics.
Fig. 14: Hillas diagram showing size and magnetic fields of potential acceleration sites. Sites below the diagonal lines cannot accelerate protons above $10^{21}$ eV, protons above $10^{21}$ eV and Fe nuclei above $10^{20}$ eV, respectively from top to bottom.

4.1 The observational case

As we alluded to above, dark matter was first identified by Fritz Zwicky [57, 58] in 1933 when studying the velocity distribution of galaxies in the Coma cluster. Using the virial theorem

$$2\langle E_{\text{kin}} \rangle = -\langle E_{\text{pot}} \rangle$$ \hspace{1cm} (155)

(where $\langle \cdots \rangle$ indicates time averaging), he concluded that there is 400 times more mass than expected from the luminosity.

This was consistently confirmed by the study of the rotation curves of galaxies i.e. the velocity $v(r)$ of stars as function of their distance $r$ to the centre of the galaxy. Using again the virial theorem (155), we have for stars in the outer regions of the galaxy, since $E_{\text{kin}} \sim mv^2$ and $E_{\text{pot}} \sim G_N m M/r$ ($m$ is the mass of the star, $M$ of the galaxy),

$$v \propto \sqrt{\frac{G_N M}{r}}.$$ \hspace{1cm} (156)

Thus, one should see the velocity decrease as $r^{-1/2}$ for stars at the border of the galaxy\textsuperscript{17}. This is not what is observed. Indeed, the rotation curves of galaxies, which were studied thoroughly through the 60s

\textsuperscript{17}Note that for stars in the interior of the galaxy one should replace $M$ by the mass of the sphere of radius $r$ ($M(r) \propto \rho_{\text{gal}} r^3$) which makes $v(r)$ increase with $r$. 

255
and 70s showed (see Figure 15) that the velocities do not start decreasing at the border of the luminous galaxy, as if there was more matter beyond. In fact, one needs a factor of order 10 more matter in spiral galaxies. The extra matter forms a halo that extends beyond the luminous galaxy.

Fig. 15: Rotation curve $v(r)$ for the galaxy NGC 3198. The curve labeled “disk” indicates the curve due to the stars in the galaxy, which extend only to 10 kpc; the curve labeled “halo” is the one that would be due solely to a spherical halo of dark matter.

For example, the modern picture of our own Galaxy, the Milky Way, is one of a luminous bulge of a few kpc at the centre of a disk of radius 12.5 kpc and thickness 0.3 kpc, containing some $10^{11}$ stars, surrounded by a nearly spherical halo of dark matter of typical radius 30 kpc (see Appendix A for the definition of a parsec).

But dark matter is not only present in galaxies. We have seen that it was first identified by Zwicky in clusters of galaxy. This is now confirmed by many observation of clusters. X-ray studies have revealed the presence of large amounts of intergalactic gas which is very hot, and hence emits X-rays. The total mass of the gas is greater than that of the galaxies by roughly a factor of two. However this is not enough mass to keep the galaxies within the cluster. Since this gas is in approximate hydrostatic equilibrium with the cluster gravitational field, the mass distribution can be determined, which leads to a total mass estimate approximately six times larger than the mass of the individual galaxies or of the hot gas.

A powerful tool for mapping dark matter is gravitational lensing, which is based on the deflection of light by matter: the light of a distant galaxy is deflected by an accumulation of matter present on the line of sight, just as it is by a lens (see Fig. 16). The deviation of light rays depends on the ratio of distances between observer, lens and source, as well as on the mass of the deflector.

More precisely, the “lens equation” may be with written as (see Fig. 15 for notations):

$$\theta_I = \theta_S + \frac{D_{LS}}{D_{OS}} \alpha.$$  \hspace{1cm} (157)

It is well-known that the deflection angle of a light ray passing an object of mass $M$ with an impact parameter $b$ is $\alpha = 4G \frac{M}{bc^2}$ (see for example [59] p.286). Hence the lensing effect allows to detect the mass distribution.
Massive clusters may induce multiple images of background galaxies. One then talks of strong lensing [60]. The shape of a single galaxy can also be deformed into an arclet through lensing. In the case of weak lensing, the effect is measured through the deformation of the shape of galaxies but, because galaxies do not have a circular shape, it can only be measured statistically: galaxies tend through lensing to have aligned shapes [61].

Finally, we have seen in Section 1.5 that cosmological data shows that dark matter is also needed at the largest scales. An illustration of this is the map of dark matter that the Planck collaboration could draw using the gravitational lensing of the CMB light (see Fig. 17).

Since the only proof of existence of dark matter is gravitational (rotation curves, lensing,...), one may wonder whether the observed phenomena are due to a modification of gravity, which would then be different from what general relativity predicts at the corresponding scales. Besides the difficulty of finding a theory that encompasses all the successes of general relativity, the problem is to modify gravity at all the scales where we see signs of dark matter, that is galaxies, clusters of galaxies and cosmological scales. For example, the MOND theory [63, 64] has been proposed to explain the rotation curves of...
galaxies but is only Newtonian and requires to be generalized [65] in order to be valid at the scale of the Universe.

The existence of the bullet cluster (see Fig. 18) where two galaxies collide has been presented as a support of dark matter [66] because the luminous parts of the galaxies are displaced with respect to their halos: because dark matter is weakly coupled, the halos (detected through gravitational lensing) continue their way during the collision whereas their luminous matter counterparts (detected through their X-ray emission) are deformed.

Fig. 18: X-ray image of the merging cluster 1E0657-558 obtained by the satellite Chandra, over which is superimposed in green contours the weak lensing reconstruction of the dark matter halos [66].

4.2 Dark matter particles

What is dark matter? Since it is nonluminous it is not electrically charged, and the only possible candidate within the Standard Model is the neutrino. But the random motion of neutrinos (the technical term is “free streaming”) would wash out any density fluctuation and prevent the formation of galaxies; one expresses this by saying that neutrinos are hot dark matter. Instead, we need cold dark matter i.e. particles with smaller free streaming length.

Moreover, we need dark matter particles in sufficient quantity. This means that they cannot be in thermal equilibrium today: they must have decoupled from the thermal history of the Universe at some early time. Typically, there are two competing effects to modify the abundance of a species $X$: $X \bar{X}$ annihilation and expansion of the Universe. Indeed, the faster is the dilution associated with the expansion, the least effective is the annihilation because the particles recede from one another. When the temperature drops below the mass $m_X$, the annihilation rate becomes smaller than the expansion rate and there is a freezing of the number of particles in a covolume. In more quantitative term, this reads for the freezing temperature $T_f$:

$$n_X(T_f) < \sigma_{\text{ann}} v > \sim H(T_f) ,$$

where $< \sigma_{\text{ann}} v >$ is the thermal average of the $X \bar{X}$ annihilation cross-section times the relative velocity of the two particles annihilating. One finds for the present density (in units of $\rho_c$, as usual) (see Ref. [67] Section 5.5)

$$\frac{\Omega_X h_0^2}{\rho_c} \approx \frac{1.07 \times 10^9 \text{ GeV}^{-1}}{g^{1/2} M_p} \frac{x_f}{< \sigma_{\text{ann}} v >} .$$

where $x_f \equiv m_X / (kT_f) \sim 20$ and $g_*$ is the total number of relativistic degrees of freedom present in the universe at the time of decoupling.
We note that the smaller the annihilation cross section is, the larger is the relic density. We find \( \Omega_X \sim (100 \text{ TeV})^{-2} \) if \( < \sigma_{\text{ann}} v > \) (in units where \( \hbar = c = 1 \), 1 pb = \( 2.5 \times 10^{-9} \) GeV\(^{-2} \)). This should be compared with the latest result \( \Omega_{DM} = 0.1187 \pm 0.0017 \) coming from Planck [16].

Thus \( \Omega_X \) will be of the right order of magnitude if \( < \sigma_{\text{ann}} v > \) is of the order of a picobarn, which is a typical order of magnitude for an electroweak process. Also writing dimensionally

\[
< \sigma_{\text{ann}} v > \sim \frac{\alpha^2}{m_X^2},
\]

where \( \alpha \) is a generic coupling strength, we find that \( \Omega_X \) is of order 1 for a mass \( m_X \sim \alpha \times 1000 \text{ TeV} \), i.e. in the TeV range. This is why one is searching for a weakly interacting massive particle (or wimp).

There is also the possibility that dark matter particles have produced non-thermally, e.g. from the decay of heavy particles.

A puzzle which is not addressed by the wimp scenario is why dark matter and baryonic matter densities are basically of the same order:

\[
\frac{\rho_{DM}}{\rho_B} \sim 5.
\]

Indeed, baryon density arises from baryogenesis (see section 1.6) and thus results from a small mismatch between baryons and antibaryons, as seen from (58). On the other hand, the wimp density results from the freezing regime described by (158). There is no reason that the two scenarios lead to similar energy densities, as in (161). This puzzle is addressed by the Asymmetric Dark Matter scenarios (see the review by C. Zurek [68] and references therein). The idea is that dark matter has an asymmetry in the number of matter over anti-matter similar to the one for baryons:

\[
N_X - N_{\bar{X}} \sim n_b - n_{\bar{b}}.
\]

The abundance is therefore approximately one part in \( 10^{10} \) in comparison with the thermal abundance (see (58)). Eq. (162) suggests that \( m_X \) is typically 5 times the proton mass, as a typical baryon mass. Thus generic Asymmetric Dark Matter models tend to favor light dark matter particles.

One may search for dark matter particles through direct detection using their elastic collisions with nuclei \( XN \rightarrow XN \) in ultra-low background detectors. The energy of the recoiling nucleus is typically from a few keV to tens of keV. The recoil rate after integration over the dark matter velocity \( v \) distribution is

\[
R \sim 3.5 \times 10^{-2} \text{ events/kg.day} \times \frac{100\text{ GeV}}{m_X} \times \frac{\sigma_{XN}}{1\text{ pb}} \times \frac{\langle v \rangle}{220\text{ km.s}^{-1}} \times \frac{\rho}{0.3\text{ GeV.cm}^{-3}}.
\]

where \( A \) is the atomic mass of the recoil nucleus, \( \sigma_{XN} \) the cross-section for dark matter particle-nucleus elastic scattering and \( \rho \) the local density of dark matter in our Galaxy. In the case of a neutralino wimp, \( \sigma_{XN} \) can be as low as \( 10^{-12} \) pb, which yields a rate of \( 10^{-8} \) events/ton.year! Present, and future, experimental limits are shown on Fig. 19. One should note that there is below \( 10^{-12} \) pb \( (10^{-8} \text{ pb for low-mass particles} \) an irreducible neutrino background corresponding to the reaction \( \nu N \rightarrow \nu N \).

Experiments of the next decade should be able to reach this limit.

Dark matter particles may also be searched for through their annihilation products in massive celestial objects. This is known as indirect detection. Indeed, because they are massive, they tend to accumulate in gravitational potentials, such as the centre of the Sun, or the centre of our Galaxy. They annihilate there into pairs of energetic particles, the energy of which is directly connected with the mass \( m_X \) of the dark matter particles. One may therefore search for excess of energetic particles (positrons,

\[\text{\[\text{This obviously precludes models where the dark matter particle is its own antiparticle, as often the case for wimps (e.g. the neutralino).}\]}}
the scalar self-coupling and $v \sim 1/(G_F \sqrt{2})^{1/2} \sim 250 \text{ GeV}$ is the Higgs vacuum expectation value. More precisely, we

WIMP–nucleon cross section \([\text{cm}^{-2}]\)

*Fig. 19:* Sensitivities of some running and planned direct detection dark matter experiments to the spin-independent elastic scattering cross-section. Full curves correspond to limits from existing experiments, dashed curves to predicted sensitivities of future experiments. The full brown, pink, blue and yellow regions correspond respectively to the regions allowed by potential signals observed by the DAMA, CREST, CDMS and CoGeNT experiments. The light red disk corresponds to the region favoured by supersymmetric models. The thick yellow line corresponds to the irreducible neutrino background.

photons, neutrinos) in the direction of galactic centres such as in our own Milky Way. An excess of energetic positrons (energy of a few tens to few hundred GeV) has actually been observed by the PAMELA, Fermi and AMS-02 experiments. It remains to be seen if this arises from the annihilation of dark matter or from astrophysical sources, such as pulsars: we have seen at the end of Section 3 that a certain number of astrophysical sources produce energetic particles. This is indeed a limitation, at least at present, of the indirect detection of dark matter: it needs to be complemented by either direct detection or detection at colliders.

### 4.3 WIMPs and physics beyond the Standard Model

We would like to stress in this section that the presence of a WIMP in a theory is deeply connected with the naturalness of the electroweak scale.

Let us start by recalling what is the naturalness problem (see for example [67]). As is well-known, the Higgs squared mass $m_H^2$ receives quadratically divergent corrections. In the context of an effective theory valid up to a cut-off scale $\Lambda$ where a more fundamental theory takes over, $\Lambda$ is the mass of the heavy degrees of freedom of the fundamental theory. Their contribution in loops, quadratic in their mass, destabilizes the Higgs mass and thus the electroweak scale ($m_H^2 \sim \lambda v^2$ where $\lambda$ is the scalar self-coupling and $v \sim 1/(G_F \sqrt{2})^{1/2} \sim 250 \text{ GeV}$ is the Higgs vacuum expectation value. More precisely, we
have at one loop
\[ \delta m_h^2 = \frac{3m_t^2\Lambda_t^2}{2\pi^2 v^2} - \frac{6M_W^2 + 3M_Z^2}{8\pi^2 v^2} \Delta_g^2 - \frac{3m_h^2\Lambda_h^2}{8\pi^2 v^2}, \]  
where for completeness we have assumed different cut-offs for the top loops (\( \Lambda_t \)), the gauge loops (\( \Lambda_g \)) and the scalar loops (\( \Lambda_h \)) [69]. The naturalness condition states that the order of magnitude of the Higgs mass is not destabilized by the radiative corrections i.e. \(|\delta m_h^2| < m_h^2\). This translates into the conditions:

\begin{align*}
\Lambda_t &\sim \frac{2}{3} \frac{\pi v}{m_t} m_h \sim 3.5 m_h, \quad (165) \\
\Lambda_g &\sim \frac{2\sqrt{2\pi v}}{\sqrt{6M_W^2 + 3M_Z^2}} m_h \sim 9 m_h, \quad (166) \\
\Lambda_h &\sim \frac{2\sqrt{2\pi v}}{\sqrt{3}} \sim 1.3 \text{ TeV}. \quad (167)
\end{align*}

Thus one should introduce new physics at a scale \( \Lambda_t \sim 3.5 m_h \). We will illustrate our argument with two examples: supersymmetry and extra dimensions. In the two cases, one introduces new physics at the scale \( \Lambda_t \) (supersymmetric particles or Kaluza-Klein modes).

Typically, these models require the presence of a symmetry that prevents direct coupling between the Standard Model (SM) fermions and the new fields that one has introduced: otherwise, such couplings introduce new mixing patterns incompatible with what is observed in flavor mixings (compatible with the Standard Model). This symmetry is usually a parity (i.e. a discrete symmetry) which is the low energy remnant of a continuous symmetry which operates at the level of the underlying fundamental theory: SM fermions are even under this parity whereas the new fields are odd. Among these new fields, the lightest odd-parity particle (we will refer to it as the LOP) is stable: it cannot decay into SM fermions because of the parity; it cannot decay into the new fields because it is the lightest. It is massive and weakly interacting. It thus provides an adequate candidate for a WIMP.

Let us take our examples in turn. In the case of supersymmetry, the parity operation is R-parity (which usually proceeds from a continuous R-symmetry broken by gaugino masses i.e. supersymmetry breaking). And the LOP is the Lightest Supersymmetric Particle, the famous LSP, the lightest neutralino in the simplest models.

In the case of extra dimensions, say a 5-dimensional model, the local symmetry is 5-dimensional Lorentz invariance. It ensures conservation of the Kaluza-Klein levels: if \( A^{(n)} \) is the \( n \)th Kaluza-Klein mode of the massless 5-dimensional field \( A \) (in other words, the 4-dimensional field with mass \( m = n/R \), where \( R \) is the radius of the 5th dimension), then in the reaction \( A^{(n)} + B^{(p)} \rightarrow C^{(q)} + D^{(r)} \), we have \( n + p = q + r \). At energies smaller than \( R^{-1} \), this turns into a Kaluza-Klein parity \((-1)^n\). The LOP is then the lightest Kaluza-Klein mode, usually \( B^{(1)} \), the first mode of the \( U(1)_{Y} \) gauge boson [70].

In realistic models, there is often the possibility that other odd-parity fields are almost degenerate in mass with the LOP. This leads to the possibility of co-annihilations, that is annihilations of the LOP against these almost degenerate fields, and to a modification of the relic density in the corresponding region of parameter space.

Searches at LHC are based on the missing energy signal corresponding to the LSP (see Fig. 20). Since LSP are produced in pairs, they are difficult to reconstruct in all generality. But, in the case of a specific model, one may be able to reconstruct the mass of the LSP as well as the relic density.

### 4.4 Other candidates for dark matter

Among the many other candidates proposed for dark matter, one may single out the axion field since it is introduced to solve one puzzle of the Standard Model known as the strong CP problem, which remains to
be solved. Moreover, from the point of view of cosmology, this is an interesting illustration of a low-mass and weakly interacting particle, as we will encounter in the next Chapter.

As we have already seen, CP symmetry is violated in weak interactions. But the symmetries of the Standard Model allow as well a CP-violating term in the QCD Lagrangian \[ (\theta g^2/32\pi^2)G_{\mu\nu}^{a}\tilde{G}^{a\mu\nu} \], where \( G_{\mu\nu} \) is the gluon field strength and \( g \) the QCD gauge coupling. For non-zero quark masses, this term leads to unobserved CP-violating effects in the strong sector\(^{19}\).

The most common way to solve the puzzle is to introduce a scalar field \( a(x) \) called axion, with Lagrangian

\[
\mathcal{L}_a = \frac{1}{2} \partial^\mu a \partial_\mu a + \frac{g^2}{32\pi^2} \frac{a(x)}{f_a} G_{\mu\nu}^{a} \tilde{G}^{a\mu\nu},
\]

(168)

where \( f_a \) is an energy scale, called the axion decay constant (by analogy with the pion decay constant). Strong interactions generate an effective potential for \( a(x) \) whose minimum corresponds to no violation of \( CP \). The non-renormalisable interaction \( aG \cdot \tilde{G} \) is obtained as the low energy effect of the breaking of a \( U(1) \) global symmetry, known as the Peccei-Quinn symmetry \[^{73, 74}\], spontaneously broken at the scale \( f_a \): the axion is the pseudo-Goldstone boson associated with this breaking \[^{75, 76}\].

---

\(^{19}\)To be more precise, it is \( \bar{\theta} \equiv \theta - \arg \det m_q \), where \( m_q \) is the quark mass matrix, which is observable. If \( \bar{\theta} \neq 0 \), then strong interactions violate \( P \) and \( CP \). This is not compatible with the experimental upper bound on the neutron electric dipole moment unless \( |\bar{\theta}| < 10^{-10} \).

\(^{20}\)I.e. a vanishing value of \( \bar{\theta} = a(x)/f_a - \arg \det m_q \).

\(^{21}\)The axion is not a true Goldstone boson but a “pseudo-Goldstone” boson because the QCD vacuum which involves non-vanishing values for the quark condensates such as \( \langle \bar{u}u \rangle \) and \( \langle \bar{d}d \rangle \) induces a small explicit breaking of the \( U(1) \) symmetry, hence a mass for the otherwise massless Goldstone boson associated with the spontaneous breaking.
The axion mass is
\[ m_a \sim 6 \times 10^{-6} \text{ eV} \left( \frac{10^{12} \text{ GeV}}{f_a} \right). \] (169)

Its coupling to ordinary matter is proportional to \(1/f_a\) and can be calculated in specific models. It couples to leptons and to photons, the latter being of the form
\[ \mathcal{L}_{a\gamma\gamma} = -g_{a\gamma} \frac{\alpha}{\pi} \frac{a(x)}{f_a} \mathbf{E} \cdot \mathbf{B}, \] (170)

where \(\mathbf{E}\) and \(\mathbf{B}\) are the electric and magnetic fields, \(\alpha\) is the fine structure constant and \(g_{a\gamma}\) is a model-dependent coefficient of order 1. Moreover, since \(m_a \ll \Lambda_{QCD}\), the axion coupling to quarks should be described through its coupling to hadrons, which occurs through small mixing with the \(\pi_0\) and \(\eta\) mesons. All of these interactions can play a role in searches for the axion, and allows the axion to be produced or detected in the laboratory and emitted by the sun or other stars. Its non-discovery leaves us with an axion window \(10^{-6} \text{ eV} < m_a < 3 \times 10^{-3} \text{ eV}\), or correspondingly, \(2.1 \times 10^{10} \text{ GeV} < f_a < 6 \times 10^{12} \text{ GeV}\).

Let us describe briefly the cosmology of the axion field (see the review by P. Sikivie [77] for a more thorough treatment). The breaking of the \(U(1)\) symmetry corresponds to a phase transition, known as the Peccei-Quinn phase transition, at a temperature of order \(f_a\). This phase transition is characterized by the formation of cosmic strings.

If the reheat temperature after inflation is smaller than \(f_a\), then one starts the evolution in the reheated universe with an homogeneous axion field. When the temperature reaches the QCD scale, the effective potential turns on and the axion acquires a mass. At a time \(t_\ast \sim m_a^{-1} (T_\ast \sim 1 \text{ GeV})\), the axion starts to oscillate around its minimum. These oscillations do not dissipate into other forms of energy, and thus contribute to the cosmological energy density an amount
\[ \Omega_a \left( \frac{h_0}{0.7} \right)^2 \sim 0.15 \left( \frac{f_a}{10^{12} \text{ GeV}} \right)^{7/6} \left( \frac{a(t_\ast)}{f_a} \right)^2, \] (171)

where \(a(t_\ast)\) is the axion value at \(t_\ast\), which measures the misalignment of the axion with respect to its final minimum. Such a contribution is thus called vacuum realignment.

If the reheat temperature is larger than \(f_a\), then the cosmic strings produced at the Peccei-Quinn phase transition become at time \(t_\ast\) the boundaries of domain walls. This leads to a potential domain wall problem (too much energy stored in the domain walls). There is a certain number of cases where this can be avoided. In this case [77],
\[ \Omega_a \left( \frac{h_0}{0.7} \right)^2 \sim 0.7 \left( \frac{f_a}{10^{12} \text{ GeV}} \right)^{7/6}. \] (172)

5 **Light does not say it all (3): dark energy**

We have seen in the introduction that the observation of the acceleration of the expansion of the Universe in 1998-1999 provided a way out of an increasingly uncomfortable tension between models and observation. On one hand, the observed (luminous and dark) matter could account for only a fraction of the critical density \(\rho_c = 10^{-26} \text{ kg/m}^3\) (other forms of energy being negligible at present time). On the other hand, the standard theory of inflation erased any curvature and naturally led to a spatially flat universe for which \(\rho = \rho_c\). The discovery of the acceleration of the expansion led to introduce a new component, named dark energy, of a type unknown so far since all known forms of energy (non-relativistic matter, radiation) decelerate the expansion.

In this last chapter, after briefly reviewing the observational case, we will illustrate dark energy models with the example of quintessence, identify some of the problems posed by dark energy models, and the fundamental questions associated with the acceleration of the expansion of the Universe.
5.1 Acceleration of the expansion of the Universe: supernovae of type Ia as standard candle

The approach that has made the first case for the acceleration of the expansion of the Universe uses supernovae of type Ia as standard candles (see Section 3.3). Two groups, the Supernova Cosmology Project [7] and the High-z Supernova Search [6] have found that distant supernovae appear to be fainter than expected in a flat matter-dominated Universe. If this is to have a cosmological origin, this means that, at fixed redshift, they are at larger distances than expected in such a context and thus that the Universe expansion is accelerating.

More precisely, one uses the relation (C.3) between the flux $\phi$ received on Earth and the luminosity $L$ of the supernova. Traditionally, flux and luminosity are expressed on a log scale as apparent magnitude $m_B$ and absolute magnitude $M$ (magnitude is $-2.5 \log_{10}$ luminosity + constant). The relation then reads

$$m_B = 5 \log(H_0d_L) + M - 5 \log H_0 + 25.$$ 

(173)

The last terms are $z$-independent, if one assumes that supernovae of type Ia are standard candles; they are then measured by using low $z$ supernovae. The first term, which involves the luminosity distance $d_L$, varies logarithmically with $z$ up to corrections which depend on the geometry, more precisely on $q_0$ for small $z$ as can be seen from (C.6). This allows to compare with data cosmological models with different components participating to the energy budget, as can be seen from Fig. 21.

![Type Ia Supernovae](image)

**Fig. 21:** Hubble plot (magnitude versus redshift) for Type Ia supernovae observed at low redshift by the Calan-Tololo Supernova Survey and at moderate redshift by the Supernova Cosmology Project.

In the case of a model with matter and cosmological constant as dominant components, $q_0 = \Omega_M/2 - \Omega_\Lambda$ and the measurement can be turned into a limit in the $\Omega_M - \Omega_\Lambda$ plane see Fig. 22).

Let us note that this combination $\Omega_M/2 - \Omega_\Lambda$ is ‘orthogonal’ to the combination $1 - \Omega_k = \Omega_M + \Omega_\Lambda$ measured in CMB experiments. The two measurements are therefore complementary: this is sometimes referred to as ‘cosmic complementarity’.

An important question raised by the analysis above is whether supernovae are truly standard candles. Otherwise, the observation could be interpreted as an history effect: for some reasons, older su-
pernovae would be dimmer. Indeed, strictly speaking, supernovae of type Ia are not standard candles: dimmer supernovae are quicker (see Section 3.3). In practice, one thus has to correct the light curves using a phenomenological stretch factor. It is thus more precise to state that supernovae of type Ia are standardizable candles. Moreover, the type of measurement discussed above is sensitive to many possible systematic effects (evolution besides the light-curve timescale correction, presence of dust, etc.), and this has fuelled a healthy debate on the significance of supernova data as well as a thorough study of possible systematic effects by the observational groups concerned.

5.2 Cosmological constant and vacuum energy

An obvious solution to the acceleration of the expansion is the introduction of a cosmological constant (see (45) remembering that $q_0$ is a deceleration parameter). But there is a severe conceptual problem associated with the cosmological constant, which we now describe.

Considering (25) in flat space at present time implies the general following constraint on $\lambda$:

$$|\lambda| \leq H_0^2.$$  

(174)

In other words, the length scale $\ell_\Lambda \equiv |\lambda|^{-1/2}$ associated with the cosmological constant must be larger than the Hubble length $\ell_{H_0} \equiv cH_0^{-1} = h_0^{-1}1.10^{26}$ m, and thus be a cosmological distance.

This is not a problem as long as one remains classical: $\ell_{H_0}$ provides a natural cosmological scale for our present Universe. The problem arises when one tries to combine gravity with the quantum theory. Indeed, from Newton’s constant and the Planck constant $\hbar$, we can construct the (reduced) Planck mass scale

$$m_P = \sqrt{\hbar c/(8\pi G_N)} = 2.4 \times 10^{18} \text{ GeV}/c^2.$$  

(175)
The corresponding length scale is the Planck length
\[ \ell_P = \frac{\hbar}{m_pc} = 8.1 \times 10^{-35} \text{ m} . \] (176)

The above constraint now reads:
\[ \ell_\Lambda \equiv |\lambda|^{-1/2} \geq \ell_{H_0} = \frac{c}{H_0} \sim 10^{60} \ell_p . \] (177)

In other words, there are more than sixty orders of magnitude between the scale associated with the cosmological constant and the scale of quantum gravity.

A rather obvious solution is to take \( \lambda = 0 \). This is as valid a choice as any other in a pure gravity theory. Unfortunately, it is an unnatural one when one introduces any kind of matter. Indeed, set \( \lambda \) to zero but assume that there is a nonvanishing vacuum (i.e. ground state) energy: \( < T_{\mu\nu} > = \rho_{\text{vac}} g_{\mu\nu} \); then the Einstein equations (3) read
\[ R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 8\pi G_N T_{\mu\nu} + 8\pi G_N \rho_{\text{vac}} g_{\mu\nu} . \] (178)

As first noted by Zel’dovich [78], the last term is interpreted as an effective cosmological constant (from now on, we set \( \hbar = c = 1 \)):
\[ \lambda_{\text{eff}} = 8\pi G_N \rho_{\text{vac}} \equiv \frac{\Lambda^4}{m_P^2} . \] (179)

Generically, \( \rho_{\text{vac}} \) receives a non-zero contribution from symmetry breaking: for instance, the scale \( \Lambda \) would be typically of the order of 100 GeV in the case of the electroweak gauge symmetry breaking or 1 TeV in the case of supersymmetry breaking. Moreover, it is divergent in the context of the (non-renormalizable) theory of gravity, which would thus favour a value as large as the Planck scale. But the constraint (177) now reads:
\[ \Lambda \leq 10^{-30} m_P \sim 10^{-3} \text{ eV} . \] (180)

It is this very unnatural fine-tuning of parameters (in explicit cases \( \rho_{\text{vac}} \) and thus \( \Lambda \) are functions of the parameters of the theory) that is referred to as the cosmological constant problem, or more accurately the vacuum energy problem.

If the acceleration observed is indeed due to the cosmological constant, its value is as large as the upper bounds obtained in the previous subsection allow:
\[ \lambda \sim H_0^2 , \quad \ell_\Lambda \sim \ell_{H_0} , \quad \Lambda \sim 10^{-3} \text{ eV} . \] (181)

Regarding the latter scale \( \Lambda \), which characterizes the vacuum energy \( (\rho_{\text{vac}} \equiv \Lambda^4) \), one may note the interesting numerical coincidence:
\[ \frac{1}{\Lambda} \sim \sqrt{\ell_{H_0} \ell_p} \sim 10^{-4} \text{ m} . \] (182)

This relation underlines the fact that the vacuum energy problem involves some deep connection between the infrared regime (the infrared cut-off being \( \ell_{H_0} \)) and the ultraviolet regime (the ultraviolet cut-off being \( \ell_p \)), between the infinitely large and the infinitely small.

As an illustration of what such a relation could tell us about some very fundamental aspects of physics (if it is not a mere numerical coincidence), we will follow the approach developed by T. Padmanabhan [79]. As we will discuss later, the value of the cosmological constant might be related to the way that the short distance theory reacts to long distance fluctuations. Let us consider a 3-dimensional domain of size \( L \) (e.g. the horizon \( \ell_{H_0} \)). From the point of view of the quantum theory, it consists of
\[ N = \left( \frac{L}{\ell_P} \right)^3 \text{ elementary cells. For each individual cell, a “natural value” for the energy stored is provided by the scale } m_p, \text{ characteristic of quantum gravity. This yields} \]
\[ \rho \sim \frac{m_p}{\ell_P^3} \sim \frac{1}{\ell_P^4}, \]  
which is some 120 orders of magnitude larger than observed.

Alternatively, if a mechanism, yet to be determined, cancels this bulk energy, vacuum energy may be produced by the energy fluctuations. The Poissonian fluctuation in energy is \[ \Delta \epsilon \sim \frac{1}{\ell_P} \]
which corresponds to an energy for overall fluctuations \[ \Delta E^2 \sim N/\ell_P^2, \] or an energy density
\[ \rho \sim \sqrt{N} \frac{1}{\ell_P L^3} \approx \frac{1}{\ell_P^2 L^2}, \]  
which again does not reproduce (182) \( (\rho_{\text{vac}} = \Lambda^4) \).

If we make the further assumption that the relevant degrees of freedom lie on the surface (as the degrees of freedom of a black hole lie on the horizon), then \[ N = \left( \frac{L}{\ell_P} \right)^2 \] and
\[ \rho \sim \frac{\sqrt{N}}{\ell_P L^3} \approx \frac{1}{\ell_P^2 L^2}, \]  
which is fully consistent with (182). Is this telling us something on the quantization of spacetime, hence of gravity? We will return to this question below.

5.3 Supersymmetry

The most natural reason why vacuum energy would be vanishing is a symmetry argument. It turns out that, among the various spacetime symmetries available, global supersymmetry is the symmetry intimately connected with the vanishing of the vacuum energy.

We recall that supersymmetry is a symmetry between bosons and fermions which plays an important rôle in high energy physics, mainly because, through a cancellation between the boson and the fermion fields, it controls severely the quantum fluctuations. Many believe that the Standard Model is the effective theory of a more fundamental theory valid at higher energies, with more boson and fermion fields. These fields should, to some level, make themselves known through the quantum energy fluctuations to which they partice. The fact that we find no trace of them seems to suggest that these fluctuations are tightly constrained, i.e. that the underlying theory is supersymmetric.

Among the various spacetime symmetries, supersymmetry is rather unique. Indeed, in the same way as the generator of time and space translations is the 4-momentum operator \[ P^\mu \equiv (P^0 = H, P^i), \] and the generator of spacetime rotations is the tensor \[ M^{\mu\nu}, \] the only other generators carrying a Lorentz index are the generators of supersymmetry \[ Q_r, \] where \( r \) is a spinor index. In fact, the combination of two supersymmetry transformations is merely a translation in spacetime. This is expressed by the following algebra:
\[ \{Q_r, \bar{Q}_s\} \gamma^0 \approx \gamma_\mu \]
(186)
where \( \bar{Q} \equiv Q \gamma^0 \) and \( \gamma^\mu \) are the gamma matrices introduced by Dirac to write a relativistic fermion (electron) equation. Since the generator of time translations \( P_t \) is the Hamiltonian \( H \), we may easily infer from (186) an expression for the Hamiltonian of the system:
\[ H = \frac{1}{4} \sum_r Q_r^2. \]  
(190)

22The anticommutator in (186) arises from the fact that the supersymmetry transformation parameter is an anticommuting spinor.
23Indeed, (186) reads explicitly
\[ \{Q_r, Q_t\} \gamma^0 = 2 \gamma^\mu P_\mu. \]  
(187)
It follows that the energy of the vacuum \(|0\rangle\) can be expressed as:
\[
\langle 0|H|0\rangle = \frac{1}{4} \sum_r \|Q_r|0\rangle\|^2 .
\] (191)

Thus, the vacuum energy vanishes if and only if supersymmetry is a symmetry of the vacuum: \(Q_r|0\rangle = 0\) for all \(r\). 24

The problem however is that, at the same time, supersymmetry predicts equal boson and fermion masses and therefore needs to be broken since this is not observed in Nature. The amount of breaking necessary to push the supersymmetric partners high enough not to have been observed yet, typically \(\Lambda \sim \text{TeV}\), is incompatible with the limit (180).

Moreover, in the context of cosmology, we should consider supersymmetry in a gravity context and thus work with its local version, supergravity (following (186), local supersymmetry transformations are associated with local spacetime translations which are nothing but the reparameterizations which play a central role in general relativity). In this context, the criterion of vanishing vacuum energy is traded for one of vanishing mass for the gravitino, the supersymmetric partner of the graviton (which allows to cancel the constant vacuum energy at the expense of generating a mass \(m_{3/2}\) for the gravitino field; see e.g. Ref. [67], section 6.3 for a more complete treatment). Local supersymmetry is then absolutely compatible with a nonvanishing vacuum energy, preferably a negative one (although possibly also a positive one). This is both a blessing and a problem: supersymmetry may be broken while the cosmological constant remains small, but we have lost our rationale for a vanishing, or very small, cosmological constant and fine-tuning raises again its ugly head.

In some supergravity theories, however, one may recover the vanishing vacuum energy criterion.

### 5.4 Why now?

In the case where the acceleration of the expansion is explained by a cosmological constant, one has to explain why this constant contribution appears to start to dominate precisely now. This is the “Why now?” or cosmic coincidence problem summarized in Fig. 23, the coincidence being between the onset of acceleration and the present time (on the scale of the age of the Universe). In order to avoid any reference to us (and hence any anthropic interpretation, see below), we may rephrase the problem as follows. Why does the dark energy starts to dominate at a time \(t_\Lambda\) (redshift \(z_\Lambda \sim 1\)) which almost coincides with the epoch \(t_G\) (redshift \(z_G \sim 3\) to \(5\)) of galaxy formation?

**Exercise 5-1**: In this exercise, we will study the evolution of a flat \((k = 0)\) universe with a non-relativistic matter component (energy density \(\rho_M\)) and a cosmological term (or equivalently vacuum energy density \(\rho_\Lambda = \lambda/(8\pi G_N)\)). The relevant equations of evolution are obtained from (23), (29) and (31):
\[
H = \frac{\dot{a}^2}{a^2} = \frac{8\pi G_N}{3}(\rho_M + \rho_\Lambda) ,
\] (192)
\[
\frac{\ddot{a}}{a} = -\frac{4\pi G_N}{3}(\rho_M - 2\rho_\Lambda) ,
\] (193)

Contracting with \(\gamma_\nu\), one obtains
\[
\sum_{r,t} \{Q_r, Q_t\} [\langle \gamma^0\rangle^2]_{rt} = 2 \text{Tr} (\gamma^0 \gamma^\nu) P_\nu .
\] (188)

Using \(\langle \gamma^0\rangle^2 = 1\) and \(\text{Tr}(\gamma^0 \gamma^\nu) = 4g^{0\nu}\), one obtains
\[
\sum_r Q_r^2 = 4P^0 = 4H .
\] (189)

24 Remember that a supersymmetry transformation \(U\) is obtained by exponentiating the generators: \(U|0\rangle = |0\rangle\).
5.5 More dynamics: dark energy vs modification of gravity

An alternate possibility is that the cosmological constant is much smaller or even vanishing and that the acceleration is due to some new form of energy –known as dark energy– or some modifications of gravity. These two possibilities correspond to modifications of either sides of Einstein’s equations (3). Let us envisage briefly these two cases.

First, we may try to identify a new component $\rho_X$ of the energy density with negative pressure:

$$ p_X = w_X \rho_X, \quad w_X < 0. $$

(197)
Note that the equation of state parameter $w_X$ may not be constant and may thus evolve with time.

Observational data constrains such a dynamical component, referred to in the literature as dark energy, just as it did with the cosmological constant. For example, in a spatially flat Universe with only matter and this unknown component $X$, one obtains from (31) with $\rho = \rho_M + \rho_X$, $p = w_X \rho_X$ the following form for the Hubble parameter and the deceleration parameter (compare with (43) and (46))

$$H^2(z) = H_0^2 \left[ \Omega_M (1 + z)^3 + \Omega_X (1 + z)^{3(1 + w_X)} \right],$$

$$q(z) = \frac{H_0^2}{2H(z)^2} \left[ \Omega_M (1 + z)^3 + \Omega_X (1 + 3w_X)(1 + z)^{3(1 + w_X)} \right],$$

where $\Omega_X = \rho_X / \rho_c$. At present time (compare with (45)),

$$q_0 = \frac{\Omega_M}{2} + (1 + 3w_X) \frac{\Omega_X}{2}.$$

The acceleration of the expansion observed requires that $\Omega_X$ dominates \(^{25}\) with $w_X < -1/3$.

In Fig. 24, we present constraints in the $(\Omega_M, w_X)$ plane, obtained recently [80] from combining observations using supernovae and BAO as well as CMB results from WMAP.

\[^{25}\text{One may easily obtain from (199) the time of the onset of the acceleration phase:}\]

$$1 + z_{acc} = \left[ (1 + 3w_X) \frac{\Omega_X}{\Omega_M} \right]^{-1/(3w_X)}.$$

**Fig. 24:** Confidence contours in the $(\Omega_M, w_X)$ plane arising from supernova results from SuperNova Legacy Survey (SNLS) 3 year results (in blue) and combined BAO/WMAP7 constraints (in green) [see Ref. [80]].

An important property of dark energy is that it does not appear to be clustered (just as a cosmological constant). Otherwise, its effects would have been detected locally, as is the case for dark matter. This points towards scalar fields, which generically have this property. Indeed, an attractive property of scalar fields is that they easily provide a diffuse background by resisting gravitational attraction. The key
quantity when discussing gravitational clustering is the speed of sound defined as

\[ c_s^2 \equiv \frac{\delta p}{\delta \rho} . \]  

(202)

It is a measure of how the pressure of the field resists gravitational clustering. In most models of dark energy, we have \( c_s^2 \sim 1 \), which explains why such scalar dark energy does not cluster: its own pressure resists gravitational collapse.

It should be stressed that, until recently, no fundamental scalar field had been observed in Nature. The discovery of a Higgs particle at the LHC high energy collider has obviously promoted the status of fundamental scalar particles.

In the next section, we will illustrate the dynamics of dark energy on the example of a scalar field evolving with time along its potential. This is often referred to as quintessence. Such scalar fields turn out to be extremely light: the only dimensionful parameter in the problem being the Hubble constant \( H_0 \), their mass is \( hH_0 \sim 10^{-33} \text{ eV} \) (see e.g. (233) below). The exchange of such fields leads to a long range force: the range is the inverse of the mass (times \( hc \)), typically \( l_{H_0} \), i.e. the size of the observable universe. This force is therefore similar to gravity and can hardly be disentangled from it. Gravitational tests such as the test of the equivalence principle thus apply not to the gravitational force alone but to the combination of gravity and this new force. This compels this force associated with dark energy to share many properties with gravity. One may thus talk of a gravitational type force.

The second possibility is to modify the left-hand side of Einstein’s equations, i.e. to look for a modification of gravity (we already mentioned that possibility in Section 4.1 for explaining observational data traditionally accounted for by dark matter). This is a notoriously difficult task because the current theory of gravity, general relativity, has passed many stringent experimental and observational tests: equivalence principle, Lorentz invariance,... Any alternative theory should first pass these tests equally successfully before being further considered.

Again, the distinction with the previous case where one introduces a new dynamical component is not as clear-cut as it would first seem. Indeed, through field redefinitions, alternate theories of gravity may be rewritten as Einstein gravity plus a dynamical scalar field. Or if one goes to more spatial dimensions, the graviton of the higher-dimensional theory may be regarded as a standard 4-dimensional graviton plus a collection of scalar (spin zero) or vector (spin one) fields.

Hence, the distinction between what is often presented as the two ways to account for the observed acceleration of the expansion is not so clear.

### 5.6 The example of quintessence

A scalar field \( \phi \) which has reached the minimum \( \phi_0 \) of its potential energy \( V(\phi) \) amounts to a cosmological constant in the form of vacuum energy: its kinetic energy is vanishing and its potential energy is the constant \( V(\phi_0) \). A more dynamical candidate for dark energy is a scalar field which is still slowly evolving in its potential [81–84]. One often refers to this field as a quintessence field.

To be more explicit, let us consider the general action which describes a real scalar field \( \phi \) minimally coupled with Einstein gravity.

\[
S = \int d^4 x \sqrt{-g} \left[ -\frac{m^2}{2} R + \frac{1}{2} \partial^\mu \phi \partial_\mu \phi - V(\phi) \right] .
\]  

(203)

Computing the corresponding energy-momentum tensor, we obtain the pressure and energy density (note the parallel with Section 2.3 where we discussed inflation models)

\[
p_{\phi} = \frac{1}{2} \dot{\phi}^2 - V(\phi) ,
\]  

(204)
\[ \rho_\phi = \frac{1}{2} \dot{\phi}^2 + V(\phi) , \]  

(205)

where, in the latter, we identify the field kinetic energy \( \dot{\phi}^2/2 \) and the potential energy \( V(\phi) \). The corresponding equation of motion is, if one neglects the spatial curvature \( (k \sim 0) \),

\[ \ddot{\phi} + 3H \dot{\phi} = -\frac{dV}{d\phi} , \]

(206)

where, besides the standard terms, one recognizes the friction term \( 3H \dot{\phi} \) due to expansion. We deduce, as expected,

\[ \rho_\phi = -3H(p_\phi + \rho_\phi) . \]

(207)

We have for the equation of state parameter

\[ w_\phi \equiv \frac{p_\phi}{\rho_\phi} = \frac{1}{2} \frac{\dot{\phi}^2 - V(\phi)}{\dot{\phi}^2 + V(\phi)} \geq -1 . \]

(208)

If the kinetic energy is subdominant \( (\dot{\phi}^2/2 \ll V(\phi)) \), we clearly obtain \(-1 \leq w_\phi \leq 0\). In any case \(-1 \leq w_\phi \leq +1\).

Let us look in more details at the dynamics of such a quintessence scalar field. For this purpose, it is useful to identify scaling solutions which we define as solutions where the \( \phi \) energy density scales as a power of the cosmic scale factor:

\[ \rho_\phi \propto a^{-n_\phi} , \quad n_\phi \text{ constant} . \]

(209)

This will allow us to identify two of the main examples of dynamical potential for quintessence:

- the exponential potential:

\[ V(\phi) = V_0 e^{-\lambda \phi} ; \]

(210)

- the Ratra-Peebles potential [82, 83],

\[ V(\phi) = \frac{M^{4+\alpha}}{\phi^\alpha} , \quad \alpha > 0 . \]

(211)

As we will see, the interest of such solutions is that they correspond to attractors in the cosmological evolution of the scalar field.

Since it follows from (209) that \( \dot{\rho_\phi}/\rho_\phi = -n_\phi H \), we deduce from (207) a relation between \( n_\phi \) and the equation of state parameter \( w_\phi \):

\[ w_\phi = \frac{n_\phi}{3} - 1 . \]

(212)

Hence the scaling solutions that we look for, exist only in epochs of the cosmological evolution where the equation of state parameter may be considered as constant (it could still be constant piecewise).

Since the dark energy (quintessence) density is expected to emerge from the background (radiation or matter) energy density, we consider the evolution of a scalar field \( \phi \) with constant parameter \( w_\phi \), during a phase dominated by a background fluid with equation of state parameter

\[ w_B = \frac{n_B}{3} - 1 \]

(213)

Following (30), we have \( a(t) \sim t^{2/n_B} \) \( (n_B = 4 \) for radiation, \( 3 \) for non-relativistic matter,...).
From (204) and (205), we obtain
\[ \dot{\phi}^2 = \frac{n_\phi}{3} \rho_\phi , \quad V(\phi) = \left(1 - \frac{n_\phi}{6}\right) \rho_\phi . \] (214)
Hence \( \dot{\phi}^2 \sim a^{-n_\phi} \) and \( \dot{\phi} \sim t^{-n_\phi/n_B} \). We thus distinguish two cases:

i) \( n_\phi = n_B \)

Clearly this implies \( w_\phi = w_B > 0 \) and the quintessence field \( \phi \) cannot be interpreted as the dark energy component. We have
\[ \phi = \phi_0 + \frac{2}{\lambda} \ln(t/t_0) , \] (215)
with \( \lambda \) constant. Then
\[ V(\phi) \sim \rho_\phi \sim a^{-n_\phi} \sim t^{-2} \sim e^{-\lambda \phi} . \] (216)
Hence, we find a scaling behavior for the exponential potential (210) in a background such that \( n_B = n_\phi \) \((w_B = w_\phi)\). The solution of the equation of motion (206) then reads
\[ \phi = \frac{1}{\lambda} \ln \left( \frac{V_0 \lambda^2}{2} \frac{n_B}{6 - n_B} t^2 \right) , \] (217)
and the energy density (205)
\[ \rho_\phi = \frac{12}{\lambda^2 n_B t^2} . \] (218)
Since \( H^2 = (\rho_B + \rho_\phi)/3 \sim [2/(n_B t)]^2 \),
\[ \frac{\rho_\phi}{\rho_B + \rho_\phi} \sim \frac{n_B}{\lambda^2} . \] (219)
Hence \( \rho_\phi/\rho_B \) tends to be constant in this scenario. One calls this property “tracking”. This is obviously compatible with our initial assumptions only if \( \lambda^2 > n_B \).

What happens if \( \lambda^2 \leq n_B \)?

It turns out that the scaling solution corresponds to a totally different regime: the scalar field is the dominant contribution to the energy density. We do not have to redo the calculation: it is identical to the previous one with the only changes \( w_B \rightarrow w_\phi \) or \( n_B \rightarrow n_\phi \) (for example, \( H^2 = (\rho_B + \rho_\phi)/3 \sim [2/(n_\phi t)]^2 \): the scalar energy density determines the evolution of the Universe). But then (219) reads \( 1 \sim n_\phi/\lambda^2 \), i.e.
\[ w_\phi = -1 + \frac{\lambda^2}{3} . \] (220)
Thus, if \( \lambda^2 < 2 \), the scalar field \( \phi \) may provide the dark energy component.

To summarize the two regimes that we have obtained for the exponential potential (210):

- if \( \lambda^2 \leq n_B \), the scaling solution has \( w_\phi = -1 + \lambda^2/3 \) and \( \rho_\phi/(\rho_B + \rho_\phi) \sim 1 \) (\( \phi \) is the dominant species).
- if \( \lambda^2 > n_B \), the scaling solution has \( w_\phi = w_B \) and \( \rho_\phi/(\rho_B + \rho_\phi) \sim n_B / \lambda^2 \) (the background energy density dominates; the scalar field energy density tracks it).

ii) \( n_\phi \neq n_B \)

Then \( \phi \sim t^{-n_B/n_\phi + 1} \) and we now have
\[ V(\phi) \sim \rho_\phi \sim a^{-n_\phi} \sim t^{-2n_\phi/n_B} \sim \phi^{-\frac{2n_B}{n_B - n_\phi}} . \] (221)
Hence, we find a scaling behaviour for the Ratra-Peebles potential (211) in a background characterized by \( n_B \neq n_\phi \) (or \( w_B \neq w_\phi \)). We have

\[
n_\phi = \frac{\alpha n_B}{\alpha + 2} \quad \text{or} \quad w_\phi = \frac{\alpha w_B - 2}{\alpha + 2}.
\]

The complete solution of the equation of motion (206) is

\[
\phi = \frac{\alpha(\alpha + 2)^2 n_B M^4 \alpha t^2}{2 [6(\alpha + 2) - n_B \alpha]} \cdot
\]

As we advertised in the beginning, these scaling solutions correspond to attractors in the cosmological evolution of the scalar field. We show in Fig. 25 the full phase diagram for the exponential potential with \( \lambda = 2 \) during matter domination \( (n_B = 3) \). The coordinates are \( x \equiv \dot{\phi}/(\sqrt{6}Hm_p^2) \) and \( y \equiv \sqrt{V}/(\sqrt{3}Hm_p^2) \). The stable (spiral) attractor is found at \( x = y = \sqrt{3}/8 \).

**Fig. 25:** Phase space diagram for the exponential potential (210) with \( \lambda = 2 \) and \( n_B = 3 \) (matter domination) \[85\]

**Exercise 5-2:** We show in this exercise that the solution (223) of the Ratra-Peebles potential (211) is also an attractor.

a) Show that a small perturbation \( \delta \phi \) satisfies the equation

\[
\delta \ddot{\phi} + \frac{6}{n_B t} \delta \dot{\phi} + \frac{2(\alpha + 1)}{n_B (\alpha + 2)^2 t^2} [6(\alpha + 2) - n_B \alpha] \delta \phi = 0.
\]

b) Look for a solution of the form \( \delta \phi \sim t^\gamma \) and express \( \gamma \) in terms of \( n_B \) and \( \alpha \).

c) Assume \( \alpha > 0 \) and standard values of \( n_B \). Show that the two solutions obtained in b) decay as \( \delta \phi \sim t^{-(6-n_B)/2n_B} \): the solution (223) is an attractor.

**Hints:**

\[
\gamma = -\frac{6 - n_B}{2n_B} \pm \sqrt{-\frac{3\alpha^2(3n_B - 2)(6 - n_B) - 12\alpha(n_b^2 - 16n_B + 12) - 4(n_b^2 - 36n_B + 36)}{2n_B (\alpha + 2)}}.
\]

\[225\]

c) The reduced discriminant of the second order polynomial in \( \alpha \) which is under the square root is simply \( 288n_b^3 > 0 \). The corresponding roots are then negative for \( 8 - \sqrt{52} \sim 0.79 < n_B < 6 \) and the
term is thus negative for $\alpha > 0$ and $n_B$ in this range. The square root contributes as an oscillating term and the two solutions corresponding to (225) decay as $\delta \phi \sim t^{-(6-n_B)/2n_B}$.

The reader may have noticed that the type of field solutions that we have found is very similar to those encountered at the onset of inflation. This is not surprising since we are introducing dynamics through the slow roll of a field down its potential. Let us look a little closer at the field evolution in the case of the Ratra-Peebles potential (211).

We have found the attractor scaling solution [82, 83] $\phi \propto a^{n_B/(2+\alpha)}$, $\rho_\phi \propto a^{-\alpha n_B/(2+\alpha)}$ in the case where the background density dominates. Thus $\rho_\phi$ decreases at a slower rate than the background density ($\rho_B \propto a^{-n_B}$) and tracks it until it becomes of the same order, at a given value $a_\alpha$. We thus have:

$$\frac{\dot{\phi}}{m_p} \sim \left( \frac{a}{a_\alpha} \right)^{n_B/(2+\alpha)},$$  \hspace{1cm} (226)

$$\frac{\dot{\rho}_\phi}{\rho_B} \sim \left( \frac{a}{a_\alpha} \right)^{2n_B/(2+\alpha)}. \hspace{1cm} (227)$$

**Exercise 5-3**: Compute the time $t_\Omega$ at which $\rho_\phi \sim \rho_M$ in terms of $M$ and $m_p$. Check that $\dot{\phi}$ at $t_\Omega$ does not depend on $M$.

Hints: $\rho_M \sim m_p^2/t^2$ and $\rho_\phi \sim M^{2(\alpha+4)/2+\alpha} t^{-2\alpha/2} \dot{a}^2/2 \rho_B^2$ give $t_\Omega \sim m_p^{\alpha+2}/M^{-\alpha+4}$.

The corresponding value for the equation of state parameter is given by (222):

$$w_\phi = -1 + \frac{\alpha (1 + w_B)}{2 + \alpha}. \hspace{1cm} (228)$$

Shortly after $\phi$ has reached for $a = a_\alpha$ a value of order $m_p$, it satisfies the usual slow roll conditions (using the notations (88) and (89) introduced in the context of inflation)

$$\epsilon \equiv \frac{1}{2} \left( \frac{m_p V'}{V} \right)^2 = (\alpha/2)(m_p/\phi)^2 < 1, \quad \eta \equiv \frac{m_p^2 V''}{V} = \alpha (\alpha + 1)(m_p/\phi)^2 < 1. \hspace{1cm} (229)$$

Therefore (228) provides a good approximation to the present value of $w_\phi$. Thus, at the end of the matter-dominated era, this field may provide the quintessence component that we are looking for.

Two features are interesting in this respect. One is that this scaling solution is reached for rather general initial conditions, i.e. whether $\rho_\phi$ starts of the same order or much smaller than the background energy density [86].

The second is the present value of $\rho$. Typically, since in this scenario $\phi$ is of order $m_p$ when the quintessence component emerges, we must choose the scale $M$ in such a way that $V(m_p) \sim \rho_c$. The constraint reads:

$$M \sim \left( H_0^2 m_p^{2+\alpha} \right)^{1/(4+\alpha)}. \hspace{1cm} (230)$$

We may note that this gives for $\alpha = 2, M \sim 10$ MeV, not such an atypical scale for high energy physics.

**Exercise 5-4**: In the case of slow roll, the equation of motion (206) simply reads $3H \dot{\phi} = -V'(\phi)$.

\(a\) Under this assumption, show that

$$\dot{\phi} = -\frac{4\pi G_N}{3} \frac{V'}{H^2} \sum_i (p_i + \rho_i), \hspace{1cm} (231)$$

where the summation is over all components of the Universe.
b) Deduce that, in the case where only matter and dark energy are nonnegligible at present time $t_0$ ($\Omega_M + \Omega_\phi = 1$),

$$\left. \frac{\ddot{\phi}}{V'} \right|_{t_0} \sim -\frac{1}{2} (1 - \Omega_\phi).$$

(232)

Hence slow roll requires that $\Omega_\phi \sim 1$.

Hints: a) Use $\dot{H} = -4\pi G_N \sum_i (p_i + \rho_i)$.

Moreover appealing, the quintessence idea is difficult to implement in the context of realistic models [87, 88]. The main problem lies in the fact that the quintessence field must be extremely weakly coupled to ordinary matter. This problem can take several forms:

- The quintessence field must be very light. If we return to our example of the Ratra-Peebles potential (211), $V''(m_p)$ provides an order of magnitude for the mass-squared of the quintessence component:

$$m_\phi \sim M \left( \frac{M}{m_p} \right)^{1+\alpha/2} \sim H_0 \sim 10^{-33} \text{ eV}. \tag{233}$$

using (230). The exchange of such a field leads to a long-range force: the range is typically $\ell_{H_0}$, the size of the presently observable Universe. Since this force has not been observed yet, this means that the field $\phi$ must be very weakly coupled to matter in order to comply with the constraints imposed on gravitational-type forces by the very stringent tests of the equivalence principle.

- The quintessence field is presently evolving with time. This may generate a time dependence of what we call the constants of Nature. Indeed, it turns out that, in modern particle theories, many of the constants of nature have a dynamical origin: they are expressed in terms of quantum fields which have settled down to their vacuum values in our present Universe. For example, in the models discussed above, it is difficult to find a symmetry that would prevent any coupling of the form

$$\beta \frac{\phi}{m_p} F^{\mu\nu} F_{\mu\nu} \tag{234}$$

to the gauge field kinetic term (or any given power of $\phi$). Since the quintessence behaviour is associated with time-dependent values of the field of order $m_p$, this would generate, in the absence of fine tuning, corrections of order one to the gauge coupling\(^\text{26}\). But the time dependence of the fine structure constant for example is very strongly constrained [89]: $|\dot{\alpha}/\alpha| < 5 \times 10^{-17} \text{ yr}^{-1}$. This yields a limit [87]:

$$|\beta| \leq 10^{-8} \frac{m_p H_0}{\langle \dot{\phi} \rangle}, \tag{235}$$

where $\langle \dot{\phi} \rangle$ is the average over the last $2 \times 10^9$ years. Let us recall [90] that the non-constancy of constants is not compatible with the principle of local position invariance (i.e. independence on the the location in time and space where a non-gravitational experiment is performed) which forms part of the Einstein’s equivalence principle. This in turn leads to violations of the universality of the weak equivalence principle [91].

- We have seen that, in the simplest models, the regime of interest is reached when the quintessence field $\phi$ value becomes larger than $m_p$. In this instance, as well as in the context of (single field) chaotic

\(^{26}\)For example, the quantum electrodynamics Lagrangian reads, with our conventions $L = -(1/4e^2) F^{\mu\nu} F_{\mu\nu}$. The extra term would lead to an effective electromagnetic coupling coupling $(1/4e^2_{\text{eff}}) = (1/4e^2) + \beta(\phi/m_p)$, hence a time-dependent fine structure constant $\alpha = e^2_{\text{eff}}/(\hbar c)$. 

\[276\]
inflation, there has been discussions whether this remains in the sub-Planck domain: strictly speaking, the answer is yes since what characterizes the Planck domain are energy densities of order $m_p^4$, whereas here $\rho_\phi$ remains much smaller because of the specific form of the potential (see e.g. (210) or (211)). It remains that, in such a context, one must take into account all non-renormalisable interactions of order $(\phi/m_p)^n$ compatible with the symmetries.

All the preceding shows that there is extreme fine tuning in the couplings of the quintessence field to ordinary matter, unless they are forbidden by some symmetry. This is somewhat reminiscent of the fine tuning associated with the cosmological constant. Let us stress however that the quintessence solution does not claim to solve the cosmological constant (vacuum energy) problem described above. Quintessence may explain the acceleration of the expansion of the Universe but has nothing to say about the cancellation of the bulk of the vacuum energy arising from quantum fluctuations.

5.7 Back to the cosmological constant

Let us conclude this Section by reviewing some of the attempts to address the problem of the cosmological constant.

5.7.1 Relaxation mechanisms

In the days where it was believed that vacuum energy was vanishing, one had to look for a mechanism to fully cancel the contribution of order $m_p^4$. One naturally advocated mechanisms that relaxed the cosmological constant to zero through the equations of motions of some dynamical fields.

For example, in the context of string models, any dimensionful parameter is expressed in terms of the fundamental string scale $M_s$ and of vacuum expectation values of scalar fields. The physics of the cosmological constant and of its relaxation to a vanishing value would then be associated with the dynamics of the corresponding scalar fields.

However, Steven Weinberg [92] has constrained the possible mechanisms for the relaxation of the cosmological constant by proving the following “no-go” theorem: it is not possible to obtain a vanishing cosmological constant as a consequence of the equations of motion of a finite number of fields. Weinberg’s no-go theorem relies on a series of assumptions: Lorentz invariance, finite number of constant fields, possibility of globally redefining these fields... All attempts to propose a relaxation mechanism have tried to avoid the conclusions of the theorem by relaxing one of these assumptions.

5.7.2 Anthropic considerations

The anthropic principle approach can be sketched as follows. We consider regions of spacetime with different values of $t_G$ (time of galaxy formation) and $t_\Lambda$, the time when the cosmological constant starts to dominate i.e. when the Universe enters a de Sitter phase of exponential expansion. Clearly galaxy formation must precede this phase otherwise no observer (similar to us) would be able to witness it. Thus $t_G \leq t_\Lambda$. On the other hand, regions with $t_\Lambda \gg t_G$ have not yet undergone any de Sitter phase of re-acceleration and are thus “phase-space suppressed” compared with regions with $t_\Lambda \sim t_G$. Hence the regions favoured have $t_\Lambda \sim t_G$ and thus $\rho_\Lambda \sim \rho_M$.

This was quantified by S. Weinberg [15, 93], who obtained the following bound:

$$\rho_\Lambda < \frac{\pi^2}{3} \rho_0 (1 + z_G)^3,$$

(236)

where $\rho_0$ is the present energy density and $z_G$ the redshift corresponding to galaxy formation. Using $z_G = 4.5$ as originally chosen by Weinberg [15], one finds $0 < \rho_\Lambda/P_M < 550$. More recent observations of a galaxy at $z = 8.6$ [94] or the existence of dwarf galaxies at $z \sim 10$ [95] give a larger anthropic
range: \[ 0 < \rho_\Lambda / \rho_M < 4000 \, . \] (237)

5.7.3 Emergent gravity

The alternative approach is to return to the origin of the vacuum energy problem. We stressed in Section 5.2 that this problem arises in the context of a quantum treatment of gravity (both \( \hbar \) and \( G_N \) are involved). At present we do not have a fully valid theory of quantum gravity. Presumably, it involves as well a quantum version of spacetime. It is probable that, just as our notion of continuous and elastic matter is only valid in a large distance approximation, our notion of continuous space and time is also only valid at large distance. Of course, it remains to be seen by which “objects” one should replace continuous space and time, for distances smaller than the Planck length, and what is the corresponding theory. In any case, space and time would be emergent notions, and probably also gravity. For what concerns us here, it could be that the solution to the vacuum problem should be searched in this deeper context. And maybe dark energy is telling us something about this underlying theory. Let us also note that, if spacetime is an emergent notion, then its symmetries are also emergent: one may expect at some level violations of Lorentz invariance for example, which lead to violations of Einstein’s equivalence principle.

5.7.4 Holography

Until now we have considered gravity as a fundamental force which is on the same footing as the other three. However, one aspect of gravity is strikingly different from what we encounter with other interactions: it is the phenomenon of gravitational collapse. As we have seen in Section 3.2, if a quantity \( E \) of (gravitating) energy is localized in a region of spacetime of size \( R \) smaller than the Schwarzschild radius defined as:

\[ R_S \equiv 2 \frac{G_N E}{c^2} \] (238)

it undergoes gravitational collapse. This has been used by some (e.g. [96]) to consider that the high-energy (ultraviolet) regime of gravity is classical: before reaching Planckian energies, regions of spacetime undergo gravitational collapse and turn into black holes, which are classical objects. This may have some far reaching consequences for the issues we are dealing with here, especially vacuum energy.

Indeed, let us return to the considerations that led to the estimate \( \rho \sim m_p / \ell_p^3 \) (see (183)) for the vacuum energy density in the context of quantum field theory. Consider a spherical region of radius \( R \) and energy density given by (183): \( \rho = m_p^4 \). Then the total energy reads:

\[ E = \frac{4 \pi}{3} R^3 \rho = \frac{4 \pi}{3} m_p (R m_p)^3 \] (239)

But the system will undergo gravitational collapse when \( R < R_S \) that is, using (238) \( R < (R m_p)^3 / 3m_p \), i.e. \( R > 1 / m_p = \ell_p \). Hence, for any volume larger than the elementary cell, on cannot concentrate a vacuum energy density \( \rho = m_p^4 \), at least in the case (that we consider here) that vacuum energy is gravitating: the system is unstable and undergoes gravitational collapse. The maximal energy density for a macroscopic region of size \( R \) is \( E < R/(2G_N) \)

\[ \rho_{\text{max}} = E / (4 \pi R^3 / 3) = 3 \frac{3}{8 \pi G_N R^2} \] (240)

Let us extend these considerations to the whole observable Universe of size \( R \sim H_0^{-1} \): we can only store a vacuum energy density

\[ \rho < \frac{3 H_0^2}{8 \pi G_N} = \rho_c \] (241)
Taken at face value, this would mean that the vacuum energy density has the value it has because our observable Universe is very large. This cannot be true at all times: otherwise, one can easily check that the presence of dark energy can be absorbed in a redefinition of Newton’s constant; up to this redefinition, the Universe would behave as if there is no dark energy and thus the recent phase of acceleration of the expansion would remain unexplained. If pushed to its full consequences, this leads to a new way of considering the quantum evolution of the Universe [97].

5.8 Concluding remarks
The most fascinating aspect of the dark energy problem is the number of fundamental questions it connects with: how did the Universe emerge from a quantum state to become so large and so old? why does dark energy emerge so late in the evolution of the Universe? does it relate to the nature of space and time as we know them? has spacetime emerged from something else? is general relativity the ultimate theory of gravity? if not, are its basic principles violated at some scale? what is a quantum state of the Universe? what is the status of an observer in such a Universe? are there multiple universes? are there more than four dimensions?...

One of the reasons is that dark energy appears to be connected with vacuum energy, which is the most fundamental issue faced by theorists in fundamental physics, an issue that illustrates the difficulties encountered at the interface between general relativity, the present theory of gravity, and the quantum theory. In some sense, the situation is reminiscent of the one encountered at the end of the XIXth century, where one had two very successful theories, Newtonian gravity and electromagnetism (summarized into the Maxwell equations). The Michelson-Morley experiment in 1887 was the experimental observation that led Einstein and others to reconsider the foundations. Similarly, both general relativity and the quantum theory, the latter described at the level of (non-gravitational) fundamental interactions by the Standard Model, are extremely successful theories. Moreover, the recent successes of cosmology have shown that our picture of the early Universe based on these two pillars is not simply qualitative but is supported quantitatively by increasingly precise observations. Is dark energy the signal of a new era? It remains to be seen its exact connection with the issue of vacuum energy. But more importantly, it is at present a conceptual difficulty, rather than a clear experimental sign of the inconsistency of the overall picture. We are still lacking our Michelson-Morley experiment.

From this perspective, it is reassuring that we have in front of us in the next decade or so a very substantial experimental programme, which includes not only increasingly precise and complete observational data, but also experiments of many types, that might help us identify the road to follow in order to reconsider the foundations of physics.

Appendices
A Astrophysical constants and scales

Constants

The tradition in astrophysics is to use the CGS system. Whereas there are in some specific cases useful quantities to be defined (such as the parsec), centimeter and gram seem hardly relevant. We thus use here the international system. Note that 1 kg.m$^{-3}$ = $10^{-3}$ g.cm$^{-3}$, 1 J = $10^7$ erg, 1 W = $10^7$ erg.s$^{-1}$.

Speed of light: $c = 299 792 458$ m s$^{-1}$

Newtonian gravitational constant: $G_N = 6.6742 \times 10^{-11}$ m$^3$ kg$^{-1}$ s$^{-2}$

$\alpha_G \equiv G_N m_p^2/(hc) = 5.906 \times 10^{-39}$

Fine structure constant: $\alpha \equiv e^2/(4\pi\epsilon_0hc) = 7.297 \times 10^{-3} = 1/137$

Thomson cross section: $\sigma_T = 8\pi r_e^2/3 = 0.665$ barn $= 0.665 \times 10^{-28}$ m$^2$
Boltzmann constant: \( k_B = 1.380 \times 10^{-23} \text{ J.K}^{-1} = 8.617 \times 10^{-5} \text{ eV.K}^{-1} \)
Planck constant: \( h = 1.054 \times 10^{-34} \text{ J.s} \)

Typical length scales

- Planck length: \( \ell_p = \sqrt{\frac{8\pi G}{c^3}} \frac{\hbar}{c} = 8.1 \times 10^{-35} \text{ m} \)
- Classical electron radius: \( r_e = \frac{e^2}{4\pi\epsilon_0mc^2} = 2.817 \times 10^{-15} \text{ m} \)
- Solar radius: \( R_\odot = 6.9598 \times 10^8 \text{ m} \)
- Astronomical unit (au) = Sun-Earth distance = 1.4960 \times 10^{11} \text{ m} 
- Parsec (au/arc sec): 1 pc = 3.262 light-year = 3.086 \times 10^{16} \text{ m} 
- Sun-galactic center distance: 10 kpc
- Milky way galaxy disk radius (luminous matter): 15 kpc
- Presently visible universe: 6 Gpc

Typical mass scales

- Reduced Planck mass: \( m_p = \sqrt{\hbar c/(8\pi G_N)} = 2.14 \times 10^{18} \text{ GeV/c}^2 = 3.81 \times 10^{-9} \text{ kg} \)
- Solar mass: \( M_\odot = 1.989 \times 10^{30} \text{ kg} \)
- Milky Way galaxy mass: 4 to 10 \times 10^{11} M_\odot

Typical luminosities

- Solar luminosity: \( L_\odot = 3.85 \times 10^{33} \text{ erg/s} = 3.85 \times 10^{26} \text{ W} \)

Typical densities

- Present mean density of the universe: \( \rho_0 \sim \rho_c = 10^{-26} \text{ kg.m}^{-3} \)
- Interstellar medium: \( 10^{-22} \text{ kg.m}^{-3} \)
- Sun: \( \rho_\odot = 1408 \text{ kg/m}^{-3} \)
- Neutron star: \( 10^{18} \text{ kg.m}^{-3} \)

**B General relativity**

In the context of general relativity, one defines the Christoffel symbol or affine connection \( \Gamma^\rho_{\mu\nu} \) which is the analogue of the gauge field (it appears in covariant derivatives). It is defined in terms of the metric as:

\[
\Gamma^\rho_{\mu\nu} = \frac{1}{2} g^{\rho\sigma} \left[ \partial_\mu g_{\nu\sigma} + \partial_\nu g_{\mu\sigma} - \partial_\sigma g_{\mu\nu} \right],
\]

where \( g^{\rho\sigma} \) is the inverse metric tensor: \( g^{\rho\sigma} g_{\sigma\tau} = \delta^\rho_\tau \).

In the same way that one defines the field strength by differentiating the gauge field, one introduces the Riemann curvature tensor:

\[
R^\nu_{\nu\alpha\beta} = \partial_\alpha \Gamma^\nu_{\beta\beta} - \partial_\beta \Gamma^\nu_{\alpha\beta} + \Gamma^\nu_{\alpha\sigma} \Gamma^\sigma_{\beta\beta} - \Gamma^\nu_{\beta\sigma} \Gamma^\sigma_{\alpha\beta}.
\]

By contracting indices, one then defines the Ricci tensor \( R_{\mu\nu} \) and the curvature scalar \( R \)

\[
R_{\mu\nu} \equiv R^\rho_{\mu\nu\rho}, \quad R \equiv g^{\mu\nu} R_{\mu\nu}.
\]
One also uses the Christoffel symbols to define the covariant derivatives:
\[ V_{\mu;\nu} = \nabla_\nu V_\mu \equiv \partial_\nu V_\mu - \Gamma^\rho_{\mu\nu} V_\rho, \]
\[ V^\mu;\nu = \nabla_\nu V^\mu \equiv \partial_\nu V^\mu + \Gamma^\mu_{\nu\rho} V^\rho. \]  

**Exercise B-1**: In the case of the Robertson-Walker metric (19),

a) compute the non-vanishing Christoffel symbols (B.1),

b) using the fact that the Ricci tensor associated with the 3-dimensional metric \( \gamma_{ij} \) is simply \( R_{ij}(\gamma) = 2k\gamma_{ij} \), compute the components of the Ricci tensor and the scalar curvature (B.3),

c) deduce the components of the Einstein tensor \( G_{\mu\nu} \) defined in (3): The components of the Einstein tensor now read (see Exercise 2-1):

\[ G_{tt} = 3 \left( \frac{\dot{a}^2}{a^2} + \frac{k}{a^2} \right), \]  
\[ G_{ij} = -\gamma_{ij} \left( \dot{a}^2 + 2a\ddot{a} + k \right), \]  

Hints: a) \( \Gamma^i_{jt} = \delta^i_j \dot{a}/a, \Gamma^i_{ij} = a\ddot{a}\gamma_{ij}, \Gamma^i_{jk} = \Gamma^i_{jk}(\gamma). \)

b) \( R_{tt} = -3\ddot{a}/a, R_{ij} = (2k + a\dddot{a} + 2\dot{a}^2) \gamma_{ij}, R = -6 \left( k + \dddot{a} + \dot{a}^2 \right)/a^2. \)

**C Measure of distances**

Measuring cosmological distances allows to study the geometry of spacetime. Depending on the type of observation, one may define several distances.

First consider a photon travelling in an expanding or contracting Friedmann universe. Its equation of motion is fixed by the condition \( ds^2 = 0 \) (as in Eq. (32) of Chapter 1). One then defines the proper distance as

\[ d(t) \equiv a(t) \int_0^r \frac{dr}{\sqrt{1 - kr^2}} = a(t) \int_0^t \frac{cdt'}{a(t')} . \]  

Using

\[ \int_0^t \frac{cdt}{a(t)} = \int_0^{a_0} \frac{cd\alpha}{\alpha\ddot{a}} = \int_0^{a_0} \frac{cd\alpha}{\alpha^2 H} = \int_0^z \frac{cdz}{H(z)} \]

we may extract from (C.1) the proper distance at time \( t_0 \):

\[ d(t_0) = a_0 \int_0^r \frac{dr}{\sqrt{1 - kr^2}} = a_0 \begin{cases} \sin^{-1} r & k = +1 \\ r & k = 0 \\ \sinh^{-1} r & k = -1 \end{cases} \]

\[ = \ell_{H_0} \int_0^z \left[ \Omega_M(1 + z)^3 + \Omega_R(1 + z)^4 + \Omega_k(1 + z)^2 + \Omega_\Lambda \right]^{1/2} \]

where \( \ell_{H_0} = cH_0^{-1} \).

If a photon source of luminosity \( L \) (energy per unit time) is placed at a distance \( r \) from the observer, then the energy flux \( \phi \) (energy per unit time and unit area) received by the observer is given by

\[ \phi = \frac{L}{4\pi a_0^2 r^2 (1 + z)^2} \equiv \frac{L}{4\pi d_L^2} . \]

The two powers of \( 1 + z \) account for the photon energy redshift and the time dilatation between emission and observation. The quantity \( d_L \equiv a_0 r (1 + z) \) is called luminosity distance.
If the source is at a redshift \( z \) of order one or smaller, the effect of spatial curvature is unimportant and we can approximate the integral \( \int_0^r dr/\sqrt{1 - kr^2} \) in (C.1) by simply \( r \) (i.e. the value for \( k = 0 \)). This equation gives

\[
\frac{a_0}{r} \sim \int_0^{a_0} \frac{a_0 c dt}{a(t)} = \int_a^{a_0} \frac{a_0 c da}{a} \sim \ell_H \int_a^{a_0} \frac{da}{a[1 - q_0 H_0(t - t_0)]} \tag{C.4}
\]

where we have used the development (47) with \( t_H = \ell_H/c = H_0^{-1} \). Using \( H_0(t - t_0) \sim (a - a_0)/a_0 \ll 1 \) and \( a = a_0/(1 + z) \), we obtain for \( z \ll 1 \)

\[
a_0 r = \ell_H z \left( 1 - \frac{1 + q_0}{2} z + \cdots \right) \tag{C.5}
\]

Thus, the luminosity distance reads, for \( z \ll 1 \),

\[
d_L = \ell_H z \left( 1 - \frac{1 + q_0}{2} z + \cdots \right) (1 + z) = \ell_H z \left( 1 + \frac{1 - q_0}{2} z + \cdots \right) \tag{C.6}
\]

Hence measurement of deviations to the Hubble law (\( d_L = \ell_H z \)) at moderate redshift allow to measure the combination \( \Omega_M/2 - \Omega_A \) (see (45)).

Another distance is defined in cases where one measures the angular diameter \( \delta \) of a source in the sky. If \( D \) is the diameter of the source, then \( D/\delta \) would be the distance of the source in Euclidean geometry. In a universe with a Robertson-Walker metric, it turns out to be \( a(t) r = a_0 r/(1 + z) \). This defines the angular diameter distance \( d_A \)

\[
d_A = \frac{d_L}{(1 + z)^2} \tag{C.7}
\]

Several distance measurements tend to point towards an evolution of the present universe dominated by the cosmological constant contribution\(^{27}\) and thus a late acceleration of its expansion, as we will now see.

Exercise C.1 : We compute exactly the luminosity distance \( d_L = a_0 r(1 + z) \) or angular distance \( d_A = a_0 r/(1 + z) \) in the case of a matter-dominated universe. Defining

\[
\zeta_k(r) \equiv \begin{cases} 
\text{sin}^{-1} r & k = +1 \\
\text{r} & k = 0 \\
\text{sinh}^{-1} r & k = -1 
\end{cases}
\tag{C.8}
\]

use (C.2) which reads, in the case of a matter-dominated universe,

\[
a_0 \zeta_k(r) = \ell_H \int_0^r \frac{dz}{[\Omega_M (1 + z)^3 + (1 - \Omega_M)(1 + z)^2]^{1/2}} \tag{C.9}
\]

to prove Mattig’s formula [98]:

\[
a_0 r = 2 \ell_H \frac{\Omega_M z + (\Omega_M - 2) \left[ \sqrt{1 + \Omega_M z} - 1 \right]}{\Omega_M^2 (1 + z)} \tag{C.10}
\]

Hints: For \( k \neq 0 \), change to the coordinate \( u^2 = k(\Omega - 1)/[\Omega (1 + z)] \) in order to compute the integral (C.9). Using the last of equations (40), which reads \( \ell_H^2 / a_0^2 = k(\Omega - 1) \), one obtains

\[
\zeta_k(r) = 2 \left( \zeta_k \left[ \sqrt{\frac{k(\Omega - 1)}{\Omega}} \right] - \zeta_k \left[ \sqrt{\frac{k(\Omega - 1)}{(1 + z)\Omega}} \right] \right),
\]

from which (C.10) can be inferred.

\(^{27}\)at least when analyzed in the framework of the model discussed in this section, i.e. including non-relativistic matter, radiation and a cosmological constant.
We study in this Appendix the perturbations of a scalar field coupled to gravity, following Ref. [99]. This has obvious implications for the study of inflation or dark energy models.

We consider the most general local action for a scalar field coupled to Einstein gravity:

\[ S = -\frac{m^2}{2} \int d^4x \sqrt{g} R + \int d^4x \sqrt{g} p(X, \phi) , \]  

(D.1)

where we have defined

\[ X \equiv \frac{1}{2} g_{\mu\nu} \partial_\mu \phi \partial_\nu \phi . \]  

(D.2)

One may describe this system as a perfect fluid, with the standard energy-momentum tensor (21):

\[ T_{\mu\nu} = -pg_{\mu\nu} + (p + \rho)U_\mu U_\nu , \]  

(D.3)

Indeed, varying with respect to the metric, we find

\[ \delta S = \int \sqrt{g} \left[ Xp_{,X} U_\mu U_\nu - \frac{1}{2} pg_{\mu\nu} \right] \delta g^{\mu\nu} \]

\[ = \frac{1}{2} \int T_{\mu\nu} \delta g^{\mu\nu} , \]  

(D.4)

with \( U_\mu \equiv \partial_\mu \phi / (2X)^{1/2} \). Thus, the energy-momentum tensor has the form (D.3) with the function \( p(X, \phi) \), i.e. the scalar Lagrangian, as the pressure (hence the notation) and the energy density:

\[ \rho = 2Xp_{,X} - p . \]  

(D.5)

In the case where \( p = X - V(\phi) \), one recovers (205,204).

In the following, a quantity will play an important role; it is the speed of sound:

\[ c^2_s \equiv \frac{\delta p}{\delta \rho} = \frac{p_{,X}}{\rho_{,X}} = \frac{p + \rho}{2Xp_{,X}} . \]  

(D.6)

We start with a background metric described by (19) (for simplicity, we assume that space is flat: \( k = 0 \); for the general case, see Ref. [99]) and with a background scalar configuration \( \varphi(t) \) which satisfies (29) (or equivalently the scalar field equation of motion).

Perturbing this background, we write in the longitudinal gauge [100]\textsuperscript{28}

\[ ds^2 = (1 + 2\Phi)dt^2 - (1 - 2\Phi)a^2(t) \delta_{ij}dx^i dx^j , \]  

(D.7)

where \( \Phi \) is the Newtonian potential, and we take for the scalar field

\[ \phi(t, x) = \varphi(t) + \delta \varphi(t, x) . \]  

(D.8)

Then

\[ \delta G_0^0 = 2 \left[ \frac{1}{a^2} \Delta \Phi - 3H \dot{\Phi} - 3H^2 \Phi \right] , \]  

(D.9)

\[ \delta G_i^0 = 2 \left[ \Phi + H \dot{\Phi} \right]_{,i} , \]  

(D.10)

\[ ^{28}\text{We use the fact that the spatial part of the energy-momentum tensor is diagonal. Otherwise, two different functions } \Phi \text{ and } \Psi \text{ would appear respectively as } (1 + 2\Phi) \text{ in the time component and } (1 - 2\Psi) \text{ in the space component [100].} \]
where, as usual, $H^2 = 8\pi G_N \rho /3$. As for the variation of the energy-momentum tensor, we have

$$\delta T^0_0 = \delta \rho = \rho,X \delta X + \rho,\phi \delta \phi \quad , \quad \delta T^0_i = (p + \rho) \delta U_i .$$

(D.11)

Note for the latter that $U_0 = 1, U_i = 0$ but $\delta U_i = (\delta \phi / \dot{\phi})_i \neq 0$. For the former, we use $\dot{\rho} = -3H(p + \rho) = \rho,X \dot{X} + \rho,\phi \dot{\phi}$. One finds

$$\delta T^0_0 = -3H(p + \rho) \left( \frac{\delta \phi}{\dot{\phi}} \right) , \quad \delta T^0_i = (p + \rho) \left( \frac{\delta \phi}{\dot{\phi}} \right)_i ,$$

(D.12)

We thus obtain from $\delta G_{\mu\nu} = 8\pi G_N \delta T_{\mu\nu}$

$$\left( \frac{\delta \phi}{\dot{\phi}} \right) = \left( 1 + \frac{c_s^2}{4\pi G_N a^2(p + \rho)} \Delta \right) \Phi ,$$

(D.14)

$$a \Phi = 4\pi G_N a(p + \rho) \left( \frac{\delta \phi}{\dot{\phi}} \right) .$$

(D.15)

The other Einstein’s equations are redundant. We may now define the new variables $\xi$ and $\zeta$:

$$a \Phi = 4\pi G_N H \xi , \quad \frac{\delta \phi}{\dot{\phi}} = \frac{\zeta}{H} = \frac{4\pi G_N}{a} \xi ,$$

(D.16)

which satisfy the equations of motion

$$\dot{\xi} = \frac{a(p + \rho)}{H^2} \zeta ,$$

(D.17)

$$\dot{\zeta} = \frac{c_s^2 H^2}{a^3(p + \rho)} \Delta \xi .$$

(D.18)

Defining

$$z \equiv \frac{a(p + \rho)^{1/2}}{c_s H}$$

(D.19)

and differentiating with respect to conformal time ($\xi' \equiv d\xi / d\eta = a\dot{\xi}$ and so on), we may write the system of differential equations simply as

$$\xi' = c_s^2 z \zeta , \quad \zeta' = \frac{1}{z^2} \Delta \xi ,$$

(D.20)

which can be turned into a single differential equation for $\zeta$. Indeed, defining $v \equiv z\zeta$, we find

$$v'' - c_s^2 \Delta v - \frac{z''}{z} v = 0 .$$

(D.21)

This can be derived from the following action:

$$S = \frac{1}{2} \int \left[ v'^2 + c_s^2 v \Delta v + \frac{z''}{z} v^2 \right] d\eta d^3x .$$

(D.22)

If we look for plane wave solutions of (D.21) i.e. $v = v_k e^{-ikx}$, we find two regimes:

- at long wavelength ($k^2 c_s^2 \ll |z''/z|$), a non-decaying solution $v_k \propto z$.
- at short wavelength ($k^2 c_s^2 \ll |z''/z|$), an oscillating solution $v_k \propto \exp(ikc_s\eta)$.
Quantization of scalar field in curved spacetime

We now turn to the quantization of the scalar degrees of freedom. The fact that we are in a non-trivial background gravitational field brings some new features but, since the background is only time dependent, the quantization procedure may be broadly inspired by the flat spacetime case.

Let us consider a generic scalar field (this could be for example the gravitational potential). The standard commutation relations

\[
[\Phi(\eta, \mathbf{x}), \Phi(\eta, \mathbf{x}')] = \left[\Pi(\eta, \mathbf{x}), \Pi(\eta, \mathbf{x}')\right] = 0 , \quad \left[\Phi(\eta, \mathbf{x}), \Pi(\eta, \mathbf{x}')\right] = i\delta^3(\mathbf{x} - \mathbf{x}') \quad (D.23)
\]

involves the canonical momentum \(\Pi = \delta L/\delta \partial_\eta \Phi\) (we are using here the conformal time \(\eta\)).

One may decompose the operator \(\Phi\) over the complete orthonormal basis of the eigenfunctions of the Laplace operator. In the spatially flat case that we are considering here, these are simply the plane waves: \(\chi_k(\eta)e^{-ik.x}\) (we note that we are making full use of spatial translation invariance, which remains a symmetry). We thus write

\[
\Phi(\eta, \mathbf{x}) = \frac{1}{\sqrt{2}} \int \frac{d^3k}{(2\pi)^{3/2}} \left[ e^{-ik.x}\chi_k(\eta)a_k^\dagger + e^{ik.x}\chi_k^*(\eta)a_k \right] , \quad (D.24)
\]

where the operators \(a_k\) and \(a_k^\dagger\) satisfy the commutation rules

\[
[a_k, a_{k'}^\dagger] = \left[ a_k^\dagger, a_{k'}^\dagger \right] = 0 , \quad \left[ a_k, a_{k'} \right] = \delta^3(\mathbf{k} - \mathbf{k'}) . \quad (D.25)
\]

This is consistent with \((D.23)\) under the condition

\[
\chi'_k(\eta)\chi_k^*(\eta) - \chi''_k(\eta)\chi_k(\eta) = 2i . \quad (D.26)
\]

The \(\chi_k(\eta)\) modes satisfy an equation of the type

\[
\chi''_k(\eta) + E_k^2\chi_k(\eta) = 0 , \quad (D.27)
\]

where \(E_k^2\) includes a mass-squared term and possibly other contributions such as the one that would arise from a non-minimal coupling of the scalar field to gravity.

In Minkowski spacetime, on constructs a Fock space of states obtained by applying a product of creation (negative frequency) operators on the vacuum state \(|0\rangle\), defined as the state annihilated by all positive frequency operators \(a_k^\dagger\): \(a_k|0\rangle = 0\). This relies on the invariance under the Poincaré group which gives an absolute meaning to these notions. More precisely, in Minkowski spacetime, the operator \(\partial/\partial t\) is a Killing vector orthogonal to the spacelike hypersurfaces \(t = \text{constant}\) and the plane wave modes \(e^{-ik.z}\) are eigenfunctions of this Killing vector with eigenvalues \(-ik_0 = -i\omega\) of a given sign.

In curved spacetime (see for example the book by Birrell and Davies [101]), the Poincaré group is no longer a symmetry group of spacetime and correspondingly there is no time-invariant notion of positive or negative frequency. There is thus no possibility of agreeing on a specific vacuum state for all inertial measuring devices. One may still rely, in some cases, on specific symmetries such as translation invariance, conformal symmetry or the de Sitter group to constrain the description of vacuum states.

[See the review by Mukhanov, Feldman and Brandenberger [100]]

We thus choose a given time \(\eta_0\) in order to define a vacuum state \(|0\rangle_{\eta_0}\) such that, for all \(k\), \(a_k|0\rangle_{\eta_0} = 0\). These annihilation operators are the operator factors of the positive frequency modes \(\chi_k^+ \equiv \chi_k(\eta)\) in the expansion \((D.24)\) (similarly we define the negative frequency modes \(\chi_k^- \equiv \chi_k(\eta)\)).

\[285\]
If all $E_k$ are positive, it turns out that one find such modes: they are the solutions of (D.27) with the following initial conditions at time $\eta_0$:

$$\chi_k(\eta_0) = E_k^{-1/2}(\eta_0), \quad \chi_k'(\eta_0) = iE_k^{1/2}(\eta_0), \quad (D.28)$$

consistent with the consistency condition (D.26). Since these solutions obviously depend on $\eta_0$, we will affect them a superscript $(0)$ in what follows.

At a later time $\eta_1$, we define along the same lines a new vacuum $|0\rangle_{\eta_1}$, which is annihilated by all operators $b_k$. These operators appear in an expansion of the type (D.24) but with new positive frequency modes $\chi_k^{(1)}$. Since equation (D.27) is linear, there is a linear relation between the positive and negative frequency modes at $\eta_0$ and $\eta_1$:

$$\chi_k^{(1)+} = \alpha_k\chi_k^{(0)+} + \beta_k\chi_k^{(0)-}, \quad |\alpha_k|^2 - |\beta_k|^2 = 1, \quad (D.29)$$

where we have used (D.26).

This defines the Bogoliubov coefficients $\alpha_k$ and $\beta_k$. Obviously we have in parallel for the operators

$$b_k = +\alpha_k a_k - \beta_k a_k^\dagger, \quad b_k^\dagger = -\beta_k a_k + \beta_k a_k^\dagger \quad (D.30)$$

Let us give an example to illustrate the physical meaning of the Bogoliubov coefficients. We start with the vacuum $|0\rangle_{\eta_0}$ at time $\eta_0$ and compute at $\eta_1$ the number of particles $b_k^\dagger b_k$. It is given by

$$\eta_0 \langle 0 | b_k^\dagger b_k | 0\rangle_{\eta_0} = |\beta_k|^2, \quad (D.31)$$

where we have used (D.30). Thus, even though we have prepared the system in the vacuum state at time $\eta_0$, the number of particles is non-vanishing at time $\eta_1$. Fluctuations can be produced quantum mechanically from the vacuum through the coupling of the scalar field to gravity.

If not all energies are positive, then we cannot define a set of modes through the boundary conditions (D.28). This is in particular the situation encountered in the case of inflation. There, the symmetries of de Sitter space help to define the so-called de Sitter invariant vacuum through the conditions:

$$\chi_k(\eta_0) = \frac{1}{k^{3/2}} (\mathcal{H}_0 + ik), \quad \chi_k'(\eta_0) = \frac{i}{k^{1/2}} (\mathcal{H}_0 + ik - i\mathcal{H}_0'/k), \quad (D.32)$$

where $\mathcal{H}_0 = a'(\eta_0)/a(\eta_0) = a(\eta_0)\mathcal{H}_0$. We recover (D.28) at small wavelength i.e. for $k \ll \mathcal{H}_0$. We note that, whereas the small wavelength behavior is universal, the large wavelength behaviour strongly depends on the choice of vacuum.

**Perturbations**

Let us now apply this formalism to the quantum generation of perturbations. We are interested in fluctuations of the Newtonian gravitational potential. Using (D.15) and (D.16), we have

$$\zeta = \Phi \left[ 1 + \frac{2}{3} \frac{\rho}{p + \rho} \right] + \frac{2}{3} \frac{\rho}{p + \rho} \frac{\dot{\Phi}}{\dot{H}}, \quad (D.33)$$

Since $\Phi$ is constant in any phase where $p/\rho$ is constant (say matter or radiation domination), then in such a phase, $\zeta$ is simply proportional to the gravitational potential. We thus consider the scalar variable $\zeta$ in what follows and write:

$$\zeta(\eta, x) = \frac{1}{\sqrt{2}} \int \frac{d^3k}{(2\pi)^{3/2}} \left[ e^{-ik.x}\zeta_k(\eta)a_k^\dagger + e^{ik.x}\zeta_k^*(\eta)a_k \right], \quad (D.34)$$
We have \( ζ_k = v_k / z \), where \( v_k \) satisfies, according to (D.21),
\[
v''_k + \left( c^2 s_k^2 - \frac{z''}{z} \right) v_k = 0 . \tag{D.35}
\]

One often characterizes the fluctuations through the power spectrum \( P_κ^ζ(η) \) which is defined from the 2-point correlation function:
\[
\langle 0 | ζ(x, η)ζ(x + r, η) | 0 \rangle = \int_{k=0}^{k=+∞} \frac{dk \sin kr}{k} P_κ^ζ(η) . \tag{D.36}
\]

Using the decomposition (D.34), one easily obtains
\[
P_κ^ζ(η) = \frac{k^3}{2π^2} |ζ_k|^2 = \frac{k^3}{2π^2} |v_k|^2 . \tag{D.37}
\]

As we have seen above, Eq. (D.35) has two distinct regimes depending of the relative magnitude of \( c_s k \) and \( z''/z \). In the case of slow roll inflation, the main dependence with time in \( z \), as given in (D.19), comes from the scale factor. Hence \( z''/z \sim a''/a \sim (aH)^2 \). Hence we have to compare \( c_s k \) with \( aH \) i.e. the comoving wavelength \( a/k \) with the sound horizon length \( c_s / H \).

In the case of short wavelength (smaller than the sound horizon), the normalized solution is
\[
v_k = \frac{1}{(2kc_s)^{1/2}} e^{ikc_s η} . \tag{D.38}
\]

For long wavelengths (larger than the sound horizon), it is
\[
v_k = C_k z , \tag{D.39}
\]
where the constant \( C_k \) may be obtained by continuity between the two approximate solutions at the scale of sound horizon: \( |C_k|^2 = 1/(2kc_s z_s) \), with \( z_s \) the value of \( z \) at horizon crossing.

We thus have the behavior indicated on Fig. 9: the quantum fluctuations are created at small wavelength and grow until they cross the (sound) horizon. From then on they grow mechanically until they renenter the horizon and are observed in the matter dominated epoch. We thus obtain the power spectrum:
\[
P_κ^ζ = \frac{k^3}{2π^2} |ζ_k|^2 \bigg|_{k=aH/c_s} = \frac{1}{c_s p + ρ} \left( \frac{H}{2π} \right)^2 \bigg|_{k=aH/c_s} . \tag{D.40}
\]

The spectral index is defined as
\[
n_S(k) - 1 = \frac{d \ln P_κ^ζ}{d \ln k} . \tag{D.41}
\]
Thus, inflation predicts a departure from a scale invariant spectrum \( (n_S = 1) \).

**Exercise D-1**: Show that, in the case of slow roll inflation (Section 2.3), the spectral index is simply given by
\[
n_S = 1 - 6ε + 2η . \tag{D.42}
\]

**Hints**: In this case, \( c_s = 1 \). Moreover,
\[
n_S(k) - 1 \sim \frac{1}{H} \frac{d \ln P_κ^ζ}{dt} = -6 \left( 1 + \frac{ρ}{H} \right) - 2 \frac{\ddot{H}}{H} \phi .
\]

Use then (90).
References


Organizing Committee

T. Donskova (JINR, Russia)
N. Ellis (CERN, Switzerland)
C. Grojean (CERN, Switzerland)
M. Mulders (CERN, Switzerland)
A. Olchevsky (JINR, Russia)

Local Organizing Committee

V. Brouillard (LAL Orsay, France)
C. Collard (IPHC Strasbourg, France)
F. Couderc (CEA Saclay, France)
P. Crochet (LPC Clermont Ferrand, France)
J-B. de Vivie de Régie (LAL Orsay, France)
S. Henrot-Versille (chair) (LAL Orsay, France)
E. Kou (LAL Orsay, France)
F. Le Diberde (LAL Orsay, France)
J. Malclès (CEA Saclay, France)
B. Viaud (LAL Orsay, France)

International Advisors

V. Matveev (JINR, Russia)
A. Skrinsky (BINP, Russia)
N. Tyurin (IHEP, Russia)

Lecturers

J. Iliopoulos (ENS Paris, France)
D. Kazakov (JINR, Russia)
G. Barenboim (University of Valencia, Spain)
L. Dixon (SLAC, USA)
P. Binétruy (APC, France)
R. Rattazzi (EPFL, Switzerland)
G. Isidori (INFN Frascati, Italy)
G. Servant (CERN, Switzerland and IPhT Saclay, France)
U. Wiedemann (CERN, Switzerland)
H. Prosper (Florida State University, USA)
G. Rolandi (CERN, Switzerland)

Discussion Leaders

A. Arbuzov (JINR, Russia)
J-B. de Vivie de Régie (LAL Orsay, France)
A. Gladychev (JINR, Russia)
G. Moreau (LPT, Orsay University, France)
M. Pierini (CERN, Switzerland)
J-C. Winter (CERN, Switzerland)
Students

Karol ADAMCZYK
Faig AHMADOV
Simon AKAR
Cristiano ALPIGIANI
Felix BACHMAIR
Lukas BAENI
Gaetano BARONE
Christian BARTH
Adam Edward BARTON
Bastian BEISCHER
Sarah BERANEK
Federico BERTOLUCCI
Bugra BILIN
Olivier BONDU
Valerio BORTOLOTTO
Simone Federico BRAZZALE
Timothy BRISTOW
Lene BRYNGEMARK
Boris BULÁNEK
Vieri CANDELISE
Driss CHARFEDDINE
Adrian CHITAN
Jesper Roy CHRISTIANSEN
Robert CLARKE
Stephen COLE
Sofia Maria CONONNI
Francesco COSTANZA
Andres Guillermo DELANNOY SOTOMAYOR
Robert EBER
Philipp ELLER
Maria EROFEEVA
Yossif ESHAQ
Petr GALLUS
Christophe GOETZMANN
Garoe GONZALEZ PARRA
Alexander GRAMOLIN
Kristian GREGersen
Miroslav HAVRANEK
Anthony HAWKINS
Claudio HELLER
Oleksii IVANYTSKYI
Brett JACKSON
Hendrik JANSEN
Konstantin KANISHCHEV
Houy KEOSHKERIAN
Aleksey KHUDYAKOV
Valentin KNUNZ
Olga KOCHEBINA
Anton KRAVCHENKO
Dean LAMBERT
Valerie LANG
Joana MACHADO MIGUENS
Caitlin MALONE
Joany MANJARRÉS RAPOS
Lukas Fritz MARTI
Homero MARTINEZ
Julien MAURER
Nickolas MC COLL
Joseph McCARTIN
Samuel Ross MEEHAN
Daniel NOONAN
Stepan OBRAZTSOV
Nathaniel ODELL
Nicola ORLANDO
James PEARCE
Mathieu PERRIN-TERRIN
Ivan PRADO LONGHI
Joe PRICE
Eftychia-Sofia PROTOPAPADAKI
Eva RIBEZL
Peter Lundgaard ROENDAHL
Itamar ROTH
Francesco RUBBO
Martin RYBAR
Itzebelt SANTOYO CASTILLO
Basil SCHNEIDER
Maud SCHWOERER
Leonid SERKIN
Artur SHAIKHIEV
Rosa SIMONIELLO
Giacomo SNIDERO
Yury STEPANENKO
Nadja STROBBE
Jared STURDY
Stephen SWEDISH
Dmitry TSIRKOV
Alan TUA
Tamara VAZQUEZ SCHROEDER
Dan VLADOIU
Xiaoxiao WANG
Christoph WASICKI
Stefan WAYAND
Adam WEBBER
Jonas WEICHERT
Jeffrey WETTER
Meng XIAO
Laura ZAMBELLI
Daniele ZANZI
### Posters

<table>
<thead>
<tr>
<th>Author</th>
<th>Poster title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faig AHMADOV</td>
<td>Search for the SM Higgs boson produced in association with a W boson with the ATLAS Detector at LHC</td>
</tr>
<tr>
<td>Simon AKAR</td>
<td>Time Dependant analysis of $B^0 \rightarrow K_s \rho^0 \gamma$ Decays in BaBar</td>
</tr>
<tr>
<td>Adam Edward BARTON</td>
<td>Time dependent angular analysis of the $B^0_s \rightarrow J/\psi \phi$ decay and the extraction of $\Delta \Gamma_s$ and weak phase of B meson in ATLAS</td>
</tr>
<tr>
<td>Simone Federico BRAZZALE</td>
<td>Estimation of background arising from fake leptons in the search for a top quark partner with the ATLAS Detector</td>
</tr>
<tr>
<td>Timothy BRISTOW</td>
<td>Neutral MSSM Higgs to di-tau Searches at the ATLAS Experiment</td>
</tr>
<tr>
<td>Sofia Maria CONSONNI</td>
<td>Standard Model $H \rightarrow \tau \tau$ search in the lepton-hadron channel at the ATLAS experiment</td>
</tr>
<tr>
<td>Francesco COSTANZA</td>
<td>Search for Supersymmetry in Final States with a Single Lepton, B-jets and Missing Transverse Energy in pp Collisions</td>
</tr>
<tr>
<td>Robert EBER</td>
<td>Radiation Hard Silicon for the CMS Tracker Upgrade</td>
</tr>
<tr>
<td>Christophe GOETZMANN</td>
<td>Tracker alignment and SUSY searches in CMS</td>
</tr>
<tr>
<td>Alexander GRAMOLIN</td>
<td>Measurement of the two-photon exchange contribution in elastic $ep$ scattering at the VEPP-3 storage ring</td>
</tr>
<tr>
<td>Oleksii IVANYTSKYI</td>
<td>The critical exponents of the QCD (tri)critical endpoint within an exactly solvable model</td>
</tr>
<tr>
<td>Houry KEOSHKERIAN</td>
<td>ATLAS Liquid Argon Calorimeter Timing Study</td>
</tr>
<tr>
<td>Olga KOCHEBINA</td>
<td>Search for New Physics in very rare D decays at LHCb</td>
</tr>
<tr>
<td>Dean LAMBERT</td>
<td>Measurement of Polarisation Amplitudes and Triple Product Asymmetries in $P \rightarrow VV$ Decays at LHCb</td>
</tr>
<tr>
<td>Joana MACHADO MIGUENS</td>
<td>Higgs boson searches with the ATLAS detector in the channel $H \rightarrow WW(*) \rightarrow l\nu l\nu$</td>
</tr>
<tr>
<td>Author</td>
<td>Poster title</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Samuel Ross MEEHAN</td>
<td>Search for Heavy Vector-like Quarks in pp Collisions at $\sqrt{s} = 7$ TeV with ATLAS</td>
</tr>
<tr>
<td>Nicola ORLANDO</td>
<td>Measurement of the cross-section for b-jets produced in association with a Z boson at $\sqrt{s} = 7$ TeV with the ATLAS detector</td>
</tr>
<tr>
<td>Mathieu PERRIN-TERRIN</td>
<td>Search for the rare decays $B^0_s \rightarrow \mu^+\mu^-$ with the LHCb Experiment</td>
</tr>
<tr>
<td>Joe PRICE</td>
<td>Search for a Standard Model Higgs boson through the $H \rightarrow ZZ$ decay channels with the ATLAS detector</td>
</tr>
<tr>
<td>Francesco RUBBO</td>
<td>Top Quark Charge Asymmetry</td>
</tr>
<tr>
<td>Leonid SERKIN</td>
<td>Kinematic fitting of ttH events using a likelihood approach</td>
</tr>
<tr>
<td>Yury STEPANENKO</td>
<td>Search for the $K_{L}^{0} \rightarrow \pi^{0}\nu\bar{\nu}$ decay at the E391a experiment</td>
</tr>
<tr>
<td>Dmitry TSIRKOV</td>
<td>Study of the $pp \rightarrow {pp}_s\pi^0$ reaction at ANKE/COSY</td>
</tr>
<tr>
<td>Alan TUA</td>
<td>Search for Gluino-mediated Sbottom Production</td>
</tr>
<tr>
<td>Tamara VÁZQUEZ SCHRÖDER</td>
<td>Neutrino weighting method for top quark mass measurement in the dilepton channel</td>
</tr>
<tr>
<td>Dan VLADOIU</td>
<td>Inclusive Zbb Cross Section Measurement in proton-proton collisions at 7 TeV with the ATLAS detector</td>
</tr>
<tr>
<td>Xiaoxiao WANG</td>
<td>Measurement of the b-tag Efficiency Using Muon pTrel Methods in tbar lep+jets Events with 5 fb$^{-1}$ of Data from the ATLAS Detector</td>
</tr>
<tr>
<td>Stefan WAYAND</td>
<td>Search for Physics Beyond the Standard Model using Multileptonic Signatures with the CMS Detector</td>
</tr>
<tr>
<td>Laura ZAMBELLI</td>
<td>Use of the NA61/SHINE data to constrain neutrino flux predictions in the T2K experiment</td>
</tr>
<tr>
<td>Daniele ZANZI</td>
<td>Physics with Tau Leptons Final States in ATLAS</td>
</tr>
</tbody>
</table>